GFPose: Learning 3D Human Pose Prior With Gradient Fields

Hai Ci, Mingdong Wu, Wentao Zhu, Xiaoxuan Ma, Hao Dong, Fangwei Zhong, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4800-4810

Learning 3D human pose prior is essential to human-centered AI. Here, we present GFPose, a versatile framework to model plausible 3D human poses for various applications. At the core of GFPose is a time-dependent score network, which estimates the gradient on each body joint and progressively denoises the perturbed 3D human pose to match a given task specification. During the denoising process, GFPose implicitly incorporates pose priors in gradients and unifies various discriminative and generative tasks in an elegant framework. Despite the simplicity, GFPose demonstrates great potential in several downstream tasks. Our experiments empirically show that 1) as a multi-hypothesis pose estimator, GFPose outperforms existing SOTAs by 20% on Human3.6M dataset. 2) as a single-hypothesis pose estimator, GFPose achieves comparable results to deterministic SOTAs, even with a vanilla backbone. 3) GFPose is able to produce diverse and realistic samples in pose denoising, completion and generation tasks.

****

CXTrack: Improving 3D Point Cloud Tracking With Contextual Information

Tian-Xing Xu, Yuan-Chen Guo, Yu-Kun Lai, Song-Hai Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1084-1093

3D single object tracking plays an essential role in many applications, such as autonomous driving. It remains a challenging problem due to the large appearance variation and the sparsity of points caused by occlusion and limited sensor capabilities. Therefore, contextual information across two consecutive frames is crucial for effective object tracking. However, points containing such useful information are often overlooked and cropped out in existing methods, leading to insufficient use of important contextual knowledge. To address this issue, we propose CXTrack, a novel transformer-based network for 3D object tracking, which exploits ConteXtual information to improve the tracking results. Specifically, we design a target-centric transformer network that directly takes point features from two consecutive frames and the previous bounding box as input to explore contextual information and implicitly propagate target cues. To achieve accurate localization for objects of all sizes, we propose a transformer-based localization head with a novel center embedding module to distinguish the target from distractors. Extensive experiments on three large-scale datasets, KITTI, nuScenes and Waymo Open Dataset, show that CXTrack achieves state-of-the-art tracking performance while running at 34 FPS.

****

Deep Frequency Filtering for Domain Generalization

Shiqi Lin, Zhizheng Zhang, Zhipeng Huang, Yan Lu, Cuiling Lan, Peng Chu, Quanzeng You, Jiang Wang, Zicheng Liu, Amey Parulkar, Viraj Navkal, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11797-11807

Improving the generalization ability of Deep Neural Networks (DNNs) is critical for their practical uses, which has been a longstanding challenge. Some theoretical studies have uncovered that DNNs have preferences for some frequency components in the learning process and indicated that this may affect the robustness of learned features. In this paper, we propose Deep Frequency Filtering (DFF) for learning domain-generalizable features, which is the first endeavour to explicitly modulate the frequency components of different transfer difficulties across domains in the latent space during training. To achieve this, we perform Fast Fourier Transform (FFT) for the feature maps at different layers, then adopt a light-weight module to learn attention masks from the frequency representations after FFT to enhance transferable components while suppressing the components not conducive to generalization. Further, we empirically compare the effectiveness of adopting different types of attention designs for implementing DFF. Extensive experiments demonstrate the effectiveness of our proposed DFF and show that applying our DFF on a plain baseline outperforms the state-of-the-art methods on diffe

rent domain generalization tasks, including close-set classification and open-se
t retrieval.
****
Frame Flexible Network
Yitian Zhang, Yue Bai, Chang Liu, Huan Wang, Sheng Li, Yun Fu; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 10504-10513
Existing video recognition algorithms always conduct different training pipeline
s for inputs with different frame numbers, which requires repetitive training op
erations and multiplying storage costs. If we evaluate the model using other fra
mes which are not used in training, we observe the performance will drop signifi
cantly (see Fig.1, which is summarized as Temporal Frequency Deviation phenomeno
n. To fix this issue, we propose a general framework, named Frame Flexible Netwo
rk (FFN), which not only enables the model to be evaluated at different frames t
o adjust its computation, but also reduces the memory costs of storing multiple
models significantly. Concretely, FFN integrates several sets of training sequen
ces, involves Multi-Frequency Alignment (MFAL) to learn temporal frequency invar
iant representations, and leverages Multi-Frequency Adaptation (MFAD) to further
 strengthen the representation abilities. Comprehensive empirical validations us
ing various architectures and popular benchmarks solidly demonstrate the effecti
veness and generalization of FFN (e.g., 7.08/5.15/2.17% performance gain at Fram
e 4/8/16 on Something-Something V1 dataset over Uniformer). Code is available at
 https://github.com/BeSpontaneous/FFN.
****
Unsupervised Cumulative Domain Adaptation for Foggy Scene Optical Flow
Hanyu Zhou, Yi Chang, Wending Yan, Luxin Yan; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9569-9578
Optical flow has achieved great success under clean scenes, but suffers from res
tricted performance under foggy scenes. To bridge the clean-to-foggy domain gap,
 the existing methods typically adopt the domain adaptation to transfer the moti
on knowledge from clean to synthetic foggy domain. However, these methods unexpe
ctedly neglect the synthetic-to-real domain gap, and thus are erroneous when app
lied to real-world scenes. To handle the practical optical flow under real foggy
 scenes, in this work, we propose a novel unsupervised cumulative domain adaptat
ion optical flow (UCDA-Flow) framework: depth-association motion adaptation and
correlation-alignment motion adaptation. Specifically, we discover that depth is
 a key ingredient to influence the optical flow: the deeper depth, the inferior
optical flow, which motivates us to design a depth-association motion adaptation
 module to bridge the clean-to-foggy domain gap. Moreover, we figure out that th
e cost volume correlation shares similar distribution of the synthetic and real
foggy images, which enlightens us to devise a correlation-alignment motion adapt
ation module to distill motion knowledge of the synthetic foggy domain to the re
al foggy domain. Note that synthetic fog is designed as the intermediate domain.
 Under this unified framework, the proposed cumulative adaptation progressively
transfers knowledge from clean scenes to real foggy scenes. Extensive experiment
s have been performed to verify the superiority of the proposed method.
****
NoisyTwins: Class-Consistent and Diverse Image Generation Through StyleGANs
Harsh Rangwani, Lavish Bansal, Kartik Sharma, Tejan Karmali, Varun Jampani, R. V
enkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2023, pp. 5987-5996
StyleGANs are at the forefront of controllable image generation as they produce
a latent space that is semantically disentangled, making it suitable for image e
diting and manipulation. However, the performance of StyleGANs severely degrades
 when trained via class-conditioning on large-scale long-tailed datasets. We fin
d that one reason for degradation is the collapse of latents for each class in t
he W latent space. With NoisyTwins, we first introduce an effective and inexpens
ive augmentation strategy for class embeddings, which then decorrelates the late
nts based on self-supervision in the W space. This decorrelation mitigates colla
pse, ensuring that our method preserves intra-class diversity with class-consist

ency in image generation. We show the effectiveness of our approach on large-sca le real-world long-tailed datasets of ImageNet-LT and iNaturalist 2019, where ou r method outperforms other methods by   19% on FID, establishing a new state-of- the-art.
****
DisCoScene: Spatially Disentangled Generative Radiance Fields for Controllable 3 D-Aware Scene Synthesis
Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Sia rohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, Sergey Tulyakov; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2023, pp. 4402-4412
Existing 3D-aware image synthesis approaches mainly focus on generating a single canonical object and show limited capacity in composing a complex scene contain ing a variety of objects. This work presents DisCoScene: a 3D-aware generative m odel for high-quality and controllable scene synthesis. The key ingredient of ou r method is a very abstract object-level representation (i.e., 3D bounding boxes without semantic annotation) as the scene layout prior, which is simple to obta in, general to describe various scene contents, and yet informative to disentang le objects and background. Moreover, it serves as an intuitive user control for scene editing. Based on such a prior, the proposed model spatially disentangles the whole scene into object-centric generative radiance fields by learning on on ly 2D images with the global-local discrimination. Our model obtains the generat ion fidelity and editing flexibility of individual objects while being able to e fficiently compose objects and the background into a complete scene. We demonstr ate state-of-the-art performance on many scene datasets, including the challengi ng Waymo outdoor dataset. Our code will be made publicly available.
****
Revisiting Self-Similarity: Structural Embedding for Image Retrieval
Seongwon Lee, Suhyeon Lee, Hongje Seong, Euntai Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23412-2 3421
Despite advances in global image representation, existing image retrieval approa ches rarely consider geometric structure during the global retrieval stage. In t his work, we revisit the conventional self-similarity descriptor from a convolut ional perspective, to encode both the visual and structural cues of the image to global image representation. Our proposed network, named Structural Embedding N etwork (SENet), captures the internal structure of the images and gradually comp resses them into dense self-similarity descriptors while learning diverse struct ures from various images. These self-similarity descriptors and original image f eatures are fused and then pooled into global embedding, so that global embeddin g can represent both geometric and visual cues of the image. Along with this nov el structural embedding, our proposed network sets new state-of-the-art performa nces on several image retrieval benchmarks, convincing its robustness to look-al ike distractors. The code and models are available: https://github.com/sungonce/ SENet.
****
Minimizing the Accumulated Trajectory Error To Improve Dataset Distillation
Jiawei Du, Yidi Jiang, Vincent Y. F. Tan, Joey Tianyi Zhou, Haizhou Li; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023, pp. 3749-3758
Model-based deep learning has achieved astounding successes due in part to the a vailability of large-scale real-world data. However, processing such massive amo unts of data comes at a considerable cost in terms of computations, storage, tra ining and the search for good neural architectures. Dataset distillation has thu s recently come to the fore. This paradigm involves distilling information from large real-world datasets into tiny and compact synthetic datasets such that pro cessing the latter yields similar performances as the former. State-of-the-art m ethods primarily rely on learning the synthetic dataset by matching the gradient s obtained during training between the real and synthetic data. However, these g radient-matching methods suffer from the accumulated trajectory error caused by

the discrepancy between the distillation and subsequent evaluation. To alleviate the adverse impact of this accumulated trajectory error, we propose a novel approach that encourages the optimization algorithm to seek a flat trajectory. We show that the weights trained on synthetic data are robust against the accumulated errors perturbations with the regularization towards the flat trajectory. Our method, called Flat Trajectory Distillation (FTD), is shown to boost the performance of gradient-matching methods by up to 4.7% on a subset of images of the ImageNet dataset with higher resolution images. We also validate the effectiveness and generalizability of our method with datasets of different resolutions and demonstrate its applicability to neural architecture search.
****

Decoupling-and-Aggregating for Image Exposure Correction

Yang Wang, Long Peng, Liang Li, Yang Cao, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18115-18124

The images captured under improper exposure conditions often suffer from contrast degradation and detail distortion. Contrast degradation will destroy the statistical properties of low-frequency components, while detail distortion will disturb the structural properties of high-frequency components, leading to the low-frequency and high-frequency components being mixed and inseparable. This will limit the statistical and structural modeling capacity for exposure correction. To address this issue, this paper proposes to decouple the contrast enhancement and detail restoration within each convolution process. It is based on the observation that, in the local regions covered by convolution kernels, the feature response of low-/high-frequency can be decoupled by addition/difference operation. To this end, we inject the addition/difference operation into the convolution process and devise a Contrast Aware (CA) unit and a Detail Aware (DA) unit to facilitate the statistical and structural regularities modeling. The proposed CA and DA can be plugged into existing CNN-based exposure correction networks to substitute the Traditional Convolution (TConv) to improve the performance. Furthermore, to maintain the computational costs of the network without changing, we aggregate two units into a single TConv kernel using structural re-parameterization. Evaluations of nine methods and five benchmark datasets demonstrate that our proposed method can comprehensively improve the performance of existing methods without introducing extra computational costs compared with the original networks. The codes will be publicly available.
****

Implicit Occupancy Flow Fields for Perception and Prediction in Self-Driving

Ben Agro, Quinlan Sykora, Sergio Casas, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1379-1388

A self-driving vehicle (SDV) must be able to perceive its surroundings and predict the future behavior of other traffic participants. Existing works either perform object detection followed by trajectory forecasting of the detected objects, or predict dense occupancy and flow grids for the whole scene. The former poses a safety concern as the number of detections needs to be kept low for efficiency reasons, sacrificing object recall. The latter is computationally expensive due to the high-dimensionality of the output grid, and suffers from the limited receptive field inherent to fully convolutional networks. Furthermore, both approaches employ many computational resources predicting areas or objects that might never be queried by the motion planner. This motivates our unified approach to perception and future prediction that implicitly represents occupancy and flow over time with a single neural network. Our method avoids unnecessary computation, as it can be directly queried by the motion planner at continuous spatio-temporal locations. Moreover, we design an architecture that overcomes the limited receptive field of previous explicit occupancy prediction methods by adding an efficient yet effective global attention mechanism. Through extensive experiments in both urban and highway settings, we demonstrate that our implicit model outperforms the current state-of-the-art. For more information, visit the project website: https://waabi.ai/research/implicito.

****
CCuantuMM: Cycle-Consistent Quantum-Hybrid Matching of Multiple Shapes
Harshil Bhatia, Edith Tretschk, Zorah Lähner, Marcel Seelbach Benkner, Michael Moeller, Christian Theobalt, Vladislav Golyanik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1296-1305
Jointly matching multiple, non-rigidly deformed 3D shapes is a challenging, NP-hard problem. A perfect matching is necessarily cycle-consistent: Following the pairwise point correspondences along several shapes must end up at the starting vertex of the original shape. Unfortunately, existing quantum shape-matching methods do not support multiple shapes and even less cycle consistency. This paper addresses the open challenges and introduces the first quantum-hybrid approach for 3D shape multi-matching; in addition, it is also cycle-consistent. Its iterative formulation is admissible to modern adiabatic quantum hardware and scales linearly with the total number of input shapes. Both these characteristics are achieved by reducing the N-shape case to a sequence of three-shape matchings, the derivation of which is our main technical contribution. Thanks to quantum annealing, high-quality solutions with low energy are retrieved for the intermediate NP-hard objectives. On benchmark datasets, the proposed approach significantly outperforms extensions to multi-shape matching of a previous quantum-hybrid two-shape matching method and is on-par with classical multi-matching methods. Our source code is available at 4dqv.mpi-inf.mpg.de/CCuantuMM/
****
TrojViT: Trojan Insertion in Vision Transformers
Mengxin Zheng, Qian Lou, Lei Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4025-4034
Vision Transformers (ViTs) have demonstrated the state-of-the-art performance in various vision-related tasks. The success of ViTs motivates adversaries to perform backdoor attacks on ViTs. Although the vulnerability of traditional CNNs to backdoor attacks is well-known, backdoor attacks on ViTs are seldom-studied. Compared to CNNs capturing pixel-wise local features by convolutions, ViTs extract global context information through patches and attentions. Naively transplanting CNN-specific backdoor attacks to ViTs yields only a low clean data accuracy and a low attack success rate. In this paper, we propose a stealth and practical ViT-specific backdoor attack TrojViT. Rather than an area-wise trigger used by CNN-specific backdoor attacks, TrojViT generates a patch-wise trigger designed to build a Trojan composed of some vulnerable bits on the parameters of a ViT stored in DRAM memory through patch salience ranking and attention-target loss. TrojViT further uses parameter distillation to reduce the bit number of the Trojan. Once the attacker inserts the Trojan into the ViT model by flipping the vulnerable bits, the ViT model still produces normal inference accuracy with benign inputs. But when the attacker embeds a trigger into an input, the ViT model is forced to classify the input to a predefined target class. We show that flipping only few vulnerable bits identified by TrojViT on a ViT model using the well-known RowHammer can transform the model into a backdoored one. We perform extensive experiments of multiple datasets on various ViT models. TrojViT can classify 99.64% of test images to a target class by flipping 345 bits on a ViT for ImageNet.
****
MarS3D: A Plug-and-Play Motion-Aware Model for Semantic Segmentation on Multi-Scan 3D Point Clouds
Jiahui Liu, Chirui Chang, Jianhui Liu, Xiaoyang Wu, Lan Ma, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9372-9381
3D semantic segmentation on multi-scan large-scale point clouds plays an important role in autonomous systems. Unlike the single-scan-based semantic segmentation task, this task requires distinguishing the motion states of points in addition to their semantic categories. However, methods designed for single-scan-based segmentation tasks perform poorly on the multi-scan task due to the lacking of an effective way to integrate temporal information. We propose MarS3D, a plug-and-play motion-aware model for semantic segmentation on multi-scan 3D point clouds. This module can be flexibly combined with single-scan models to allow them to

have multi-scan perception abilities. The model encompasses two key designs: the Cross-Frame Feature Embedding module for enriching representation learning and the Motion-Aware Feature Learning module for enhancing motion awareness. Extensive experiments show that MarS3D can improve the performance of the baseline model by a large margin. The code is available at https://github.com/CVMI-Lab/MarS3D.

****

An Image Quality Assessment Dataset for Portraits

Nicolas Chahine, Stefania Calarasanu, Davide Garcia-Civiero, Théo Cayla, Sira Ferradans, Jean Ponce; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9968-9978

Year after year, the demand for ever-better smartphone photos continues to grow, in particular in the domain of portrait photography. Manufacturers thus use perceptual quality criteria throughout the development of smartphone cameras. This costly procedure can be partially replaced by automated learning-based methods for image quality assessment (IQA). Due to its subjective nature, it is necessary to estimate and guarantee the consistency of the IQA process, a characteristic lacking in the mean opinion scores (MOS) widely used for crowdsourcing IQA. In addition, existing blind IQA (BIQA) datasets pay little attention to the difficulty of cross-content assessment, which may degrade the quality of annotations. This paper introduces PIQ23, a portrait-specific IQA dataset of 5116 images of 50 predefined scenarios acquired by 100 smartphones, covering a high variety of brands, models, and use cases. The dataset includes individuals of various genders and ethnicities who have given explicit and informed consent for their photographs to be used in public research. It is annotated by pairwise comparisons (PWC) collected from over 30 image quality experts for three image attributes: face detail preservation, face target exposure, and overall image quality. An in-depth statistical analysis of these annotations allows us to evaluate their consistency over PIQ23. Finally, we show through an extensive comparison with existing baselines that semantic information (image context) can be used to improve IQA predictions.

****

MSeg3D: Multi-Modal 3D Semantic Segmentation for Autonomous Driving

Jiale Li, Hang Dai, Hao Han, Yong Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21694-21704

LiDAR and camera are two modalities available for 3D semantic segmentation in autonomous driving. The popular LiDAR-only methods severely suffer from inferior segmentation on small and distant objects due to insufficient laser points, while the robust multi-modal solution is under-explored, where we investigate three crucial inherent difficulties: modality heterogeneity, limited sensor field of view intersection, and multi-modal data augmentation. We propose a multi-modal 3D semantic segmentation model (MSeg3D) with joint intra-modal feature extraction and inter-modal feature fusion to mitigate the modality heterogeneity. The multi-modal fusion in MSeg3D consists of geometry-based feature fusion GF-Phase, cross-modal feature completion, and semantic-based feature fusion SF-Phase on all visible points. The multi-modal data augmentation is reinvigorated by applying asymmetric transformations on LiDAR point cloud and multi-camera images individually, which benefits the model training with diversified augmentation transformations. MSeg3D achieves state-of-the-art results on nuScenes, Waymo, and SemanticKITTI datasets. Under the malfunctioning multi-camera input and the multi-frame point clouds input, MSeg3D still shows robustness and improves the LiDAR-only baseline. Our code is publicly available at https://github.com/jialeli1/lidarseg3d.

****

Robust Outlier Rejection for 3D Registration With Variational Bayes

Haobo Jiang, Zheng Dang, Zhen Wei, Jin Xie, Jian Yang, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1148-1157

Learning-based outlier (mismatched correspondence) rejection for robust 3D registration generally formulates the outlier removal as an inlier/outlier classification problem. The core for this to be successful is to learn the discriminative

inlier/outlier feature representations. In this paper, we develop a novel variat ional non-local network-based outlier rejection framework for robust alignment. By reformulating the non-local feature learning with variational Bayesian infere nce, the Bayesian-driven long-range dependencies can be modeled to aggregate dis criminative geometric context information for inlier/outlier distinction. Specif ically, to achieve such Bayesian-driven contextual dependencies, each query/key/ value component in our non-local network predicts a prior feature distribution a nd a posterior one. Embedded with the inlier/outlier label, the posterior featur e distribution is label-dependent and discriminative. Thus, pushing the prior to be close to the discriminative posterior in the training step enables the featu res sampled from this prior at test time to model high-quality long-range depend encies. Notably, to achieve effective posterior feature guidance, a specific pro babilistic graphical model is designed over our non-local model, which lets us d erive a variational low bound as our optimization objective for model training. Finally, we propose a voting-based inlier searching strategy to cluster the high -quality hypothetical inliers for transformation estimation. Extensive experimen ts on 3DMatch, 3DLoMatch, and KITTI datasets verify the effectiveness of our met hod.
****
Dynamically Instance-Guided Adaptation: A Backward-Free Approach for Test-Time D omain Adaptive Semantic Segmentation
Wei Wang, Zhun Zhong, Weijie Wang, Xi Chen, Charles Ling, Boyu Wang, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2023, pp. 24090-24099
In this paper, we study the application of Test-time domain adaptation in semant ic segmentation (TTDA-Seg) where both efficiency and effectiveness are crucial. Existing methods either have low efficiency (e.g., backward optimization) or ign ore semantic adaptation (e.g., distribution alignment). Besides, they would suff er from the accumulated errors caused by unstable optimization and abnormal dist ributions. To solve these problems, we propose a novel backward-free approach fo r TTDA-Seg, called Dynamically Instance-Guided Adaptation (DIGA). Our principle is utilizing each instance to dynamically guide its own adaptation in a non-para metric way, which avoids the error accumulation issue and expensive optimizing c ost. Specifically, DIGA is composed of a distribution adaptation module (DAM) an d a semantic adaptation module (SAM), enabling us to jointly adapt the model in two indispensable aspects. DAM mixes the instance and source BN statistics to en courage the model to capture robust representation. SAM combines the historical prototypes with instance-level prototypes to adjust semantic predictions, which can be associated with the parametric classifier to mutually benefit the final r esults. Extensive experiments evaluated on five target domains demonstrate the e ffectiveness and efficiency of the proposed method. Our DIGA establishes new sta te-of-the-art performance in TTDA-Seg.
****
Painting 3D Nature in 2D: View Synthesis of Natural Scenes From a Single Semanti c Mask
Shangzhan Zhang, Sida Peng, Tianrun Chen, Linzhan Mou, Haotong Lin, Kaicheng Yu, Yiyi Liao, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2023, pp. 8518-8528
We introduce a novel approach that takes a single semantic mask as input to synt hesize multi-view consistent color images of natural scenes, trained with a coll ection of single images from the Internet. Prior works on 3D-aware image synthes is either require multi-view supervision or learning category-level prior for sp ecific classes of objects, which are inapplicable to natural scenes. Our key ide a to solve this challenge is to use a semantic field as the intermediate represe ntation, which is easier to reconstruct from an input semantic mask and then tra nslated to a radiance field with the assistance of off-the-shelf semantic image synthesis models. Experiments show that our method outperforms baseline methods and produces photorealistic and multi-view consistent videos of a variety of nat ural scenes. The project website is https://zju3dv.github.io/paintingnature/.
****

LANIT: Language-Driven Image-to-Image Translation for Unlabeled Data
Jihye Park, Sunwoo Kim, Soohyun Kim, Seokju Cho, Jaejun Yoo, Youngjung Uh, Seung ryong Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23401-23411
Existing techniques for image-to-image translation commonly have suffered from two critical problems: heavy reliance on per-sample domain annotation and/or inability to handle multiple attributes per image. Recent truly-unsupervised methods adopt clustering approaches to easily provide per-sample one-hot domain labels. However, they cannot account for the real-world setting: one sample may have multiple attributes. In addition, the semantics of the clusters are not easily coupled to human understanding. To overcome these, we present LANguage-driven Image-to-image Translation model, dubbed LANIT. We leverage easy-to-obtain candidate attributes given in texts for a dataset: the similarity between images and attributes indicates per-sample domain labels. This formulation naturally enables multi-hot labels so that users can specify the target domain with a set of attributes in language. To account for the case that the initial prompts are inaccurate, we also present prompt learning. We further present domain regularization loss that enforces translated images to be mapped to the corresponding domain. Experiments on several standard benchmarks demonstrate that LANIT achieves comparable or superior performance to existing models. The code is available at github.com/KU-CVLAB/LANIT.
****
MoLo: Motion-Augmented Long-Short Contrastive Learning for Few-Shot Action Recognition
Xiang Wang, Shiwei Zhang, Zhiwu Qing, Changxin Gao, Yingya Zhang, Deli Zhao, Nong Sang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18011-18021
Current state-of-the-art approaches for few-shot action recognition achieve promising performance by conducting frame-level matching on learned visual features. However, they generally suffer from two limitations: i) the matching procedure between local frames tends to be inaccurate due to the lack of guidance to force long-range temporal perception; ii) explicit motion learning is usually ignored, leading to partial information loss. To address these issues, we develop a Motion-augmented Long-short Contrastive Learning (MoLo) method that contains two crucial components, including a long-short contrastive objective and a motion autodecoder. Specifically, the long-short contrastive objective is to endow local frame features with long-form temporal awareness by maximizing their agreement with the global token of videos belonging to the same class. The motion autodecoder is a lightweight architecture to reconstruct pixel motions from the differential features, which explicitly embeds the network with motion dynamics. By this means, MoLo can simultaneously learn long-range temporal context and motion cues for comprehensive few-shot matching. To demonstrate the effectiveness, we evaluate MoLo on five standard benchmarks, and the results show that MoLo favorably outperforms recent advanced methods. The source code is available at https://github.com/alibaba-mmai-research/MoLo.
****
Fast Point Cloud Generation With Straight Flows
Lemeng Wu, Dilin Wang, Chengyue Gong, Xingchao Liu, Yunyang Xiong, Rakesh Ranjan, Raghuraman Krishnamoorthi, Vikas Chandra, Qiang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9445-9454
Diffusion models have emerged as a powerful tool for point cloud generation. A key component that drives the impressive performance for generating high-quality samples from noise is iteratively denoise for thousands of steps. While beneficial, the complexity of learning steps has limited its applications to many 3D real-world. To address this limitation, we propose Point Straight Flow (PSF), a model that exhibits impressive performance using one step. Our idea is based on the reformulation of the standard diffusion model, which optimizes the curvy learning trajectory into a straight path. Further, we develop a distillation strategy to shorten the straight path into one step without a performance loss, enabling

applications to 3D real-world with latency constraints. We perform evaluations on multiple 3D tasks and find that our PSF performs comparably to the standard diffusion model, outperforming other efficient 3D point cloud generation methods. On real-world applications such as point cloud completion and training-free text-guided generation in a low-latency setup, PSF performs favorably.
****
Text-Guided Unsupervised Latent Transformation for Multi-Attribute Image Manipulation
Xiwen Wei, Zhen Xu, Cheng Liu, Si Wu, Zhiwen Yu, Hau San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19285-19294
Great progress has been made in StyleGAN-based image editing. To associate with preset attributes, most existing approaches focus on supervised learning for semantically meaningful latent space traversal directions, and each manipulation step is typically determined for an individual attribute. To address this limitation, we propose a Text-guided Unsupervised StyleGAN Latent Transformation (TUSLT) model, which adaptively infers a single transformation step in the latent space of StyleGAN to simultaneously manipulate multiple attributes on a given input image. Specifically, we adopt a two-stage architecture for a latent mapping network to break down the transformation process into two manageable steps. Our network first learns a diverse set of semantic directions tailored to an input image, and later nonlinearly fuses the ones associated with the target attributes to infer a residual vector. The resulting tightly interlinked two-stage architecture delivers the flexibility to handle diverse attribute combinations. By leveraging the cross-modal text-image representation of CLIP, we can perform pseudo annotations based on the semantic similarity between preset attribute text descriptions and training images, and further jointly train an auxiliary attribute classifier with the latent mapping network to provide semantic guidance. We perform extensive experiments to demonstrate that the adopted strategies contribute to the superior performance of TUSLT.
****
Achieving a Better Stability-Plasticity Trade-Off via Auxiliary Networks in Continual Learning
Sanghwan Kim, Lorenzo Noci, Antonio Orvieto, Thomas Hofmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11930-11939
In contrast to the natural capabilities of humans to learn new tasks in a sequential fashion, neural networks are known to suffer from catastrophic forgetting, where the model's performances on old tasks drop dramatically after being optimized for a new task. Since then, the continual learning (CL) community has proposed several solutions aiming to equip the neural network with the ability to learn the current task (plasticity) while still achieving high accuracy on the previous tasks (stability). Despite remarkable improvements, the plasticity-stability trade-off is still far from being solved, and its underlying mechanism is poorly understood. In this work, we propose Auxiliary Network Continual Learning (ANCL), a novel method that applies an additional auxiliary network which promotes plasticity to the continually learned model which mainly focuses on stability. More concretely, the proposed framework materializes in a regularizer that naturally interpolates between plasticity and stability, surpassing strong baselines on task incremental and class incremental scenarios. Through extensive analyses on ANCL solutions, we identify some essential principles beneath the stability-plasticity trade-off.
****
Power Bundle Adjustment for Large-Scale 3D Reconstruction
Simon Weber, Nikolaus Demmel, Tin Chon Chan, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 281-289
We introduce Power Bundle Adjustment as an expansion type algorithm for solving large-scale bundle adjustment problems. It is based on the power series expansion of the inverse Schur complement and constitutes a new family of solvers that w

e call inverse expansion methods. We theoretically justify the use of power series and we prove the convergence of our approach. Using the real-world BAL dataset we show that the proposed solver challenges the state-of-the-art iterative methods and significantly accelerates the solution of the normal equation, even for reaching a very high accuracy. This easy-to-implement solver can also complement a recently presented distributed bundle adjustment framework. We demonstrate that employing the proposed Power Bundle Adjustment as a sub-problem solver significantly improves speed and accuracy of the distributed optimization.
****
Picture That Sketch: Photorealistic Image Generation From Abstract Sketches
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), 2023, pp. 6850-6861
Given an abstract, deformed, ordinary sketch from untrained amateurs like you and me, this paper turns it into a photorealistic image - just like those shown in Fig. 1(a), all non-cherry-picked. We differ significantly from prior art in that we do not dictate an edgemap-like sketch to start with, but aim to work with abstract free-hand human sketches. In doing so, we essentially democratise the sketch-to-photo pipeline, "picturing" a sketch regardless of how good you sketch. Our contribution at the outset is a decoupled encoder-decoder training paradigm, where the decoder is a StyleGAN trained on photos only. This importantly ensures that generated results are always photorealistic. The rest is then all centred around how best to deal with the abstraction gap between sketch and photo. For that, we propose an autoregressive sketch mapper trained on sketch-photo pairs that maps a sketch to the StyleGAN latent space. We further introduce specific designs to tackle the abstract nature of human sketches, including a fine-grained discriminative loss on the back of a trained sketch-photo retrieval model, and a partial-aware sketch augmentation strategy. Finally, we showcase a few downstream tasks our generation model enables, amongst them is showing how fine-grained sketch-based image retrieval, a well-studied problem in the sketch community, can be reduced to an image (generated) to image retrieval task, surpassing state-of-the-arts. We put forward generated results in the supplementary for everyone to scrutinise. Project page: https://subhadeepkoley.github.io/PictureThatSketch
****
Contrastive Semi-Supervised Learning for Underwater Image Restoration via Reliable Bank
Shirui Huang, Keyan Wang, Huan Liu, Jun Chen, Yunsong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18145-18155
Despite the remarkable achievement of recent underwater image restoration techniques, the lack of labeled data has become a major hurdle for further progress. In this work, we propose a mean-teacher based Semi-supervised Underwater Image Restoration (Semi-UIR) framework to incorporate the unlabeled data into network training. However, the naive mean-teacher method suffers from two main problems: (1) The consistency loss used in training might become ineffective when the teacher's prediction is wrong. (2) Using L1 distance may cause the network to overfit wrong labels, resulting in confirmation bias. To address the above problems, we first introduce a reliable bank to store the "best-ever" outputs as pseudo ground truth. To assess the quality of outputs, we conduct an empirical analysis based on the monotonicity property to select the most trustworthy NR-IQA method. Besides, in view of the confirmation bias problem, we incorporate contrastive regularization to prevent the overfitting on wrong labels. Experimental results on both full-reference and non-reference underwater benchmarks demonstrate that our algorithm has obvious improvement over SOTA methods quantitatively and qualitatively. Code has been released at https://github.com/Huang-ShiRui/Semi-UIR.
****
Video Event Restoration Based on Keyframes for Video Anomaly Detection
Zhiwei Yang, Jing Liu, Zhaoyang Wu, Peng Wu, Xiaotao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14592-14601

Video anomaly detection (VAD) is a significant computer vision problem. Existing
deep neural network (DNN) based VAD methods mostly follow the route of frame re
construction or frame prediction. However, the lack of mining and learning of hi
gher-level visual features and temporal context relationships in videos limits t
he further performance of these two approaches. Inspired by video codec theory,
we introduce a brand-new VAD paradigm to break through these limitations: First,
we propose a new task of video event restoration based on keyframes. Encouragin
g DNN to infer missing multiple frames based on video keyframes so as to restore
a video event, which can more effectively motivate DNN to mine and learn potent
ial higher-level visual features and comprehensive temporal context relationship
s in the video. To this end, we propose a novel U-shaped Swin Transformer Networ
k with Dual Skip Connections (USTN-DSC) for video event restoration, where a cro
ss-attention and a temporal upsampling residual skip connection are introduced t
o further assist in restoring complex static and dynamic motion object features
in the video. In addition, we propose a simple and effective adjacent frame diff
erence loss to constrain the motion consistency of the video sequence. Extensive
experiments on benchmarks demonstrate that USTN-DSC outperforms most existing m
ethods, validating the effectiveness of our method.
****
EcoTTA: Memory-Efficient Continual Test-Time Adaptation via Self-Distilled Regul
arization
Junha Song, Jungsoo Lee, In So Kweon, Sungha Choi; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11920-119
29
This paper presents a simple yet effective approach that improves continual test
-time adaptation (TTA) in a memory-efficient manner. TTA may primarily be conduc
ted on edge devices with limited memory, so reducing memory is crucial but has b
een overlooked in previous TTA studies. In addition, long-term adaptation often
leads to catastrophic forgetting and error accumulation, which hinders applying
TTA in real-world deployments. Our approach consists of two components to addres
s these issues. First, we present lightweight meta networks that can adapt the f
rozen original networks to the target domain. This novel architecture minimizes
memory consumption by decreasing the size of intermediate activations required f
or backpropagation. Second, our novel self-distilled regularization controls the
output of the meta networks not to deviate significantly from the output of the
frozen original networks, thereby preserving well-trained knowledge from the so
urce domain. Without additional memory, this regularization prevents error accum
ulation and catastrophic forgetting, resulting in stable performance even in lon
g-term test-time adaptation. We demonstrate that our simple yet effective strate
gy outperforms other state-of-the-art methods on various benchmarks for image cl
assification and semantic segmentation tasks. Notably, our proposed method with
ResNet-50 and WideResNet-40 takes 86% and 80% less memory than the recent state-
of-the-art method, CoTTA.
****
3D-Aware Object Goal Navigation via Simultaneous Exploration and Identification
Jiazhao Zhang, Liu Dai, Fanpeng Meng, Qingnan Fan, Xuelin Chen, Kai Xu, He Wang;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2023, pp. 6672-6682
Object goal navigation (ObjectNav) in unseen environments is a fundamental task
for Embodied AI. Agents in existing works learn ObjectNav policies based on 2D m
aps, scene graphs, or image sequences. Considering this task happens in 3D space
, a 3D-aware agent can advance its ObjectNav capability via learning from fine-g
rained spatial information. However, leveraging 3D scene representation can be p
rohibitively unpractical for policy learning in this floor-level task, due to lo
w sample efficiency and expensive computational cost. In this work, we propose a
framework for the challenging 3D-aware ObjectNav based on two straightforward s
ub-policies. The two sub-polices, namely corner-guided exploration policy and ca
tegory-aware identification policy, simultaneously perform by utilizing online f
used 3D points as observation. Through extensive experiments, we show that this
framework can dramatically improve the performance in ObjectNav through learning

from 3D scene representation. Our framework achieves the best performance among all modular-based methods on the Matterport3D and Gibson datasets while requiring (up to30x) less computational cost for training. The code will be released to benefit the community.
****
Tri-Perspective View for Vision-Based 3D Semantic Occupancy Prediction
Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9223-9232
Modern methods for vision-centric autonomous driving perception widely adopt the bird's-eye-view (BEV) representation to describe a 3D scene. Despite its better efficiency than voxel representation, it has difficulty describing the fine-grained 3D structure of a scene with a single plane. To address this, we propose a tri-perspective view (TPV) representation which accompanies BEV with two additional perpendicular planes. We model each point in the 3D space by summing its projected features on the three planes. To lift image features to the 3D TPV space, we further propose a transformer-based TPV encoder (TPVFormer) to obtain the TPV features effectively. We employ the attention mechanism to aggregate the image features corresponding to each query in each TPV plane. Experiments show that our model trained with sparse supervision effectively predicts the semantic occupancy for all voxels. We demonstrate for the first time that using only camera inputs can achieve comparable performance with LiDAR-based methods on the LiDAR segmentation task on nuScenes. Code: https://github.com/wzzheng/TPVFormer.
****
Castling-ViT: Compressing Self-Attention via Switching Towards Linear-Angular Attention at Vision Transformer Inference
Haoran You, Yunyang Xiong, Xiaoliang Dai, Bichen Wu, Peizhao Zhang, Haoqi Fan, Peter Vajda, Yingyan (Celine) Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14431-14442
Vision Transformers (ViTs) have shown impressive performance but still require a high computation cost as compared to convolutional neural networks (CNNs), one reason is that ViTs' attention measures global similarities and thus has a quadratic complexity with the number of input tokens. Existing efficient ViTs adopt local attention or linear attention, which sacrifice ViTs' capabilities of capturing either global or local context. In this work, we ask an important research question: Can ViTs learn both global and local context while being more efficient during inference? To this end, we propose a framework called Castling-ViT, which trains ViTs using both linear-angular attention and masked softmax-based quadratic attention, but then switches to having only linear-angular attention during inference. Our Castling-ViT leverages angular kernels to measure the similarities between queries and keys via spectral angles. And we further simplify it with two techniques: (1) a novel linear-angular attention mechanism: we decompose the angular kernels into linear terms and high-order residuals, and only keep the linear terms; and (2) we adopt two parameterized modules to approximate high-order residuals: a depthwise convolution and an auxiliary masked softmax attention to help learn global and local information, where the masks for softmax attention are regularized to gradually become zeros and thus incur no overhead during inference. Extensive experiments validate the effectiveness of our Castling-ViT, e.g., achieving up to a 1.8% higher accuracy or 40% MACs reduction on classification and 1.2 higher mAP on detection under comparable FLOPs, as compared to ViTs with vanilla softmax-based attentions. Project page is available at https://www.haoranyou.com/castling-vit.
****
Shape, Pose, and Appearance From a Single Image via Bootstrapped Radiance Field Inversion
Dario Pavllo, David Joseph Tan, Marie-Julie Rakotosaona, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4391-4401
Neural Radiance Fields (NeRF) coupled with GANs represent a promising direction in the area of 3D reconstruction from a single view, owing to their ability to e

fficiently model arbitrary topologies. Recent work in this area, however, has mostly focused on synthetic datasets where exact ground-truth poses are known, and has overlooked pose estimation, which is important for certain downstream applications such as augmented reality (AR) and robotics. We introduce a principled end-to-end reconstruction framework for natural images, where accurate ground-truth poses are not available. Our approach recovers an SDF-parameterized 3D shape, pose, and appearance from a single image of an object, without exploiting multiple views during training. More specifically, we leverage an unconditional 3D-aware generator, to which we apply a hybrid inversion scheme where a model produces a first guess of the solution which is then refined via optimization. Our framework can de-render an image in as few as 10 steps, enabling its use in practical scenarios. We demonstrate state-of-the-art results on a variety of real and synthetic benchmarks.
****
Unlearnable Clusters: Towards Label-Agnostic Unlearnable Examples
Jiaming Zhang, Xingjun Ma, Qi Yi, Jitao Sang, Yu-Gang Jiang, Yaowei Wang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3984-3993
There is a growing interest in developing unlearnable examples (UEs) against visual privacy leaks on the Internet. UEs are training samples added with invisible but unlearnable noise, which have been found can prevent unauthorized training of machine learning models. UEs typically are generated via a bilevel optimization framework with a surrogate model to remove (minimize) errors from the original samples, and then applied to protect the data against unknown target models. However, existing UE generation methods all rely on an ideal assumption called labelconsistency, where the hackers and protectors are assumed to hold the same label for a given sample. In this work, we propose and promote a more practical label-agnostic setting, where the hackers may exploit the protected data quite differently from the protectors. E.g., a m-class unlearnable dataset held by the protector may be exploited by the hacker as a n-class dataset. Existing UE generation methods are rendered ineffective in this challenging setting. To tackle this challenge, we present a novel technique called Unlearnable Clusters (UCs) to generate label-agnostic unlearnable examples with cluster-wise perturbations. Furthermore, we propose to leverage Vision-and-Language Pretrained Models (VLPMs) like CLIP as the surrogate model to improve the transferability of the crafted UCs to diverse domains. We empirically verify the effectiveness of our proposed approach under a variety of settings with different datasets, target models, and even commercial platforms Microsoft Azure and Baidu PaddlePaddle. Code is available at https://github.com/jiamingzhang94/ Unlearnable-Clusters.
****
Rethinking Federated Learning With Domain Shift: A Prototype View
Wenke Huang, Mang Ye, Zekun Shi, He Li, Bo Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16312-16322
Federated learning shows a bright promise as a privacy-preserving collaborative learning technique. However, prevalent solutions mainly focus on all private data sampled from the same domain. An important challenge is that when distributed data are derived from diverse domains. The private model presents degenerative performance on other domains (with domain shift). Therefore, we expect that the global model optimized after the federated learning process stably provides generalizability performance on multiple domains. In this paper, we propose Federated Prototypes Learning (FPL) for federated learning under domain shift. The core idea is to construct cluster prototypes and unbiased prototypes, providing fruitful domain knowledge and a fair convergent target. On the one hand, we pull the sample embedding closer to cluster prototypes belonging to the same semantics than cluster prototypes from distinct classes. On the other hand, we introduce consistency regularization to align the local instance with the respective unbiased prototype. Empirical results on Digits and Office Caltech tasks demonstrate the effectiveness of the proposed solution and the efficiency of crucial modules.
****
NoPe-NeRF: Optimising Neural Radiance Field With No Pose Prior

Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4160-4169
Training a Neural Radiance Field (NeRF) without pre-computed camera poses is challenging. Recent advances in this direction demonstrate the possibility of jointly optimising a NeRF and camera poses in forward-facing scenes. However, these methods still face difficulties during dramatic camera movement. We tackle this challenging problem by incorporating undistorted monocular depth priors. These priors are generated by correcting scale and shift parameters during training, with which we are then able to constrain the relative poses between consecutive frames. This constraint is achieved using our proposed novel loss functions. Experiments on real-world indoor and outdoor scenes show that our method can handle challenging camera trajectories and outperforms existing methods in terms of novel view rendering quality and pose estimation accuracy. Our project page is https://nope-nerf.active.vision.
****
HGFormer: Hierarchical Grouping Transformer for Domain Generalized Semantic Segmentation
Jian Ding, Nan Xue, Gui-Song Xia, Bernt Schiele, Dengxin Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15413-15423
Current semantic segmentation models have achieved great success under the independent and identically distributed (i.i.d.) condition. However, in real-world applications, test data might come from a different domain than training data. Therefore, it is important to improve model robustness against domain differences. This work studies semantic segmentation under the domain generalization setting, where a model is trained only on the source domain and tested on the unseen target domain. Existing works show that Vision Transformers are more robust than CNNs and show that this is related to the visual grouping property of self-attention. In this work, we propose a novel hierarchical grouping transformer (HGFormer) to explicitly group pixels to form part-level masks and then whole-level masks. The masks at different scales aim to segment out both parts and a whole of classes. HGFormer combines mask classification results at both scales for class label prediction. We assemble multiple interesting cross-domain settings by using seven public semantic segmentation datasets. Experiments show that HGFormer yields more robust semantic segmentation results than per-pixel classification methods and flat-grouping transformers, and outperforms previous methods significantly. Code will be available at https://github.com/dingjiansw101/HGFormer.
****
Distilling Vision-Language Pre-Training To Collaborate With Weakly-Supervised Temporal Action Localization
Chen Ju, Kunhao Zheng, Jinxiang Liu, Peisen Zhao, Ya Zhang, Jianlong Chang, Qi Tian, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14751-14762
Weakly-supervised temporal action localization (WTAL) learns to detect and classify action instances with only category labels. Most methods widely adopt the off-the-shelf Classification-Based Pre-training (CBP) to generate video features for action localization. However, the different optimization objectives between classification and localization, make temporally localized results suffer from the serious incomplete issue. To tackle this issue without additional annotations, this paper considers to distill free action knowledge from Vision-Language Pre-training (VLP), as we surprisingly observe that the localization results of vanilla VLP have an over-complete issue, which is just complementary to the CBP results. To fuse such complementarity, we propose a novel distillation-collaboration framework with two branches acting as CBP and VLP respectively. The framework is optimized through a dual-branch alternate training strategy. Specifically, during the B step, we distill the confident background pseudo-labels from the CBP branch; while during the F step, the confident foreground pseudo-labels are distilled from the VLP branch. As a result, the dual-branch complementarity is effectively fused to promote one strong alliance. Extensive experiments and ablation s

tudies on THUMOS14 and ActivityNet1.2 reveal that our method significantly outpe rforms state-of-the-art methods.
****
Augmentation Matters: A Simple-Yet-Effective Approach to Semi-Supervised Semanti c Segmentation
Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, Jingdong Wang; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023, pp. 11350-11359
Recent studies on semi-supervised semantic segmentation (SSS) have seen fast pro gress. Despite their promising performance, current state-of-the-art methods ten d to increasingly complex designs at the cost of introducing more network compon ents and additional training procedures. Differently, in this work, we follow a standard teacher-student framework and propose AugSeg, a simple and clean approa ch that focuses mainly on data perturbations to boost the SSS performance. We ar gue that various data augmentations should be adjusted to better adapt to the se mi-supervised scenarios instead of directly applying these techniques from super vised learning. Specifically, we adopt a simplified intensity-based augmentation that selects a random number of data transformations with uniformly sampling di stortion strengths from a continuous space. Based on the estimated confidence of the model on different unlabeled samples, we also randomly inject labelled info rmation to augment the unlabeled samples in an adaptive manner. Without bells an d whistles, our simple AugSeg can readily achieve new state-of-the-art performan ce on SSS benchmarks under different partition protocols.
****
SIEDOB: Semantic Image Editing by Disentangling Object and Background
Wuyang Luo, Su Yang, Xinjian Zhang, Weishan Zhang; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1868-1878
Semantic image editing provides users with a flexible tool to modify a given ima ge guided by a corresponding segmentation map. In this task, the features of the foreground objects and the backgrounds are quite different. However, all previo us methods handle backgrounds and objects as a whole using a monolithic model. C onsequently, they remain limited in processing content-rich images and suffer fr om generating unrealistic objects and texture-inconsistent backgrounds. To addre ss this issue, we propose a novel paradigm, Semantic Image Editing by Disentangl ing Object and Background (SIEDOB), the core idea of which is to explicitly leve rages several heterogeneous subnetworks for objects and backgrounds. First, SIED OB disassembles the edited input into background regions and instance-level obje cts. Then, we feed them into the dedicated generators. Finally, all synthesized parts are embedded in their original locations and utilize a fusion network to o btain a harmonized result. Moreover, to produce high-quality edited images, we p ropose some innovative designs, including Semantic-Aware Self-Propagation Module , Boundary-Anchored Patch Discriminator, and Style-Diversity Object Generator, a nd integrate them into SIEDOB. We conduct extensive experiments on Cityscapes an d ADE20K-Room datasets and exhibit that our method remarkably outperforms the ba selines, especially in synthesizing realistic and diverse objects and texture-co nsistent backgrounds.
****
Multiclass Confidence and Localization Calibration for Object Detection
Bimsara Pathiraja, Malitha Gunawardhana, Muhammad Haris Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp . 19734-19743
Albeit achieving high predictive accuracy across many challenging computer visio n problems, recent studies suggest that deep neural networks (DNNs) tend to make overconfident predictions, rendering them poorly calibrated. Most of the existi ng attempts for improving DNN calibration are limited to classification tasks an d restricted to calibrating in-domain predictions. Surprisingly, very little to no attempts have been made in studying the calibration of object detection metho ds, which occupy a pivotal space in vision-based security-sensitive, and safety-critical applications. In this paper, we propose a new train-time technique for calibrating modern object detection methods. It is capable of jointly calibratin

g multiclass confidence and box localization by leveraging their predictive uncertainties. We perform extensive experiments on several in-domain and out-of-domain detection benchmarks. Results demonstrate that our proposed train-time calibration method consistently outperforms several baselines in reducing calibration error for both in-domain and out-of-domain predictions. Our code and models are available at https://github.com/bimsarapathiraja/MCCL
****

Query-Dependent Video Representation for Moment Retrieval and Highlight Detection
WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23023-23033
Recently, video moment retrieval and highlight detection (MR/HD) are being spotlighted as the demand for video understanding is drastically increased. The key objective of MR/HD is to localize the moment and estimate clip-wise accordance level, i.e., saliency score, to the given text query. Although the recent transformer-based models brought some advances, we found that these methods do not fully exploit the information of a given query. For example, the relevance between text query and video contents is sometimes neglected when predicting the moment and its saliency. To tackle this issue, we introduce Query-Dependent DETR (QD-DETR), a detection transformer tailored for MR/HD. As we observe the insignificant role of a given query in transformer architectures, our encoding module starts with cross-attention layers to explicitly inject the context of text query into video representation. Then, to enhance the model's capability of exploiting the query information, we manipulate the video-query pairs to produce irrelevant pairs. Such negative (irrelevant) video-query pairs are trained to yield low saliency scores, which in turn, encourages the model to estimate precise accordance between query-video pairs. Lastly, we present an input-adaptive saliency predictor which adaptively defines the criterion of saliency scores for the given video-query pairs. Our extensive studies verify the importance of building the query-dependent representation for MR/HD. Specifically, QD-DETR outperforms state-of-the-art methods on QVHighlights, TVSum, and Charades-STA datasets. Codes are available at github.com/wjun0830/QD-DETR.
****

Robust 3D Shape Classification via Non-Local Graph Attention Network
Shengwei Qin, Zhong Li, Ligang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5374-5383
We introduce a non-local graph attention network (NLGAT), which generates a novel global descriptor through two sub-networks for robust 3D shape classification. In the first sub-network, we capture the global relationships between points (i.e., point-point features) by designing a global relationship network (GRN). In the second sub-network, we enhance the local features with a geometric shape attention map obtained from a global structure network (GSN). To keep rotation invariant and extract more information from sparse point clouds, all sub-networks use the Gram matrices with different dimensions as input for working with robust classification. Additionally, GRN effectively preserves the low-frequency features and improves the classification results. Experimental results on various datasets exhibit that the classification effect of the NLGAT model is better than other state-of-the-art models. Especially, in the case of sparse point clouds (64 points) with noise under arbitrary SO(3) rotation, the classification result (85.4%) of NLGAT is improved by 39.4% compared with the best development of other methods.
****

Boosting Verified Training for Robust Image Classifications via Abstraction
Zhaodi Zhang, Zhiyi Xue, Yang Chen, Si Liu, Yueling Zhang, Jing Liu, Min Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16251-16260
This paper proposes a novel, abstraction-based, certified training method for robust image classifiers. Via abstraction, all perturbed images are mapped into intervals before feeding into neural networks for training. By training on interva

ls, all the perturbed images that are mapped to the same interval are classified as the same label, rendering the variance of training sets to be small and the loss landscape of the models to be smooth. Consequently, our approach significantly improves the robustness of trained models. For the abstraction, our training method also enables a sound and complete black-box verification approach, which is orthogonal and scalable to arbitrary types of neural networks regardless of their sizes and architectures. We evaluate our method on a wide range of benchmarks in different scales. The experimental results show that our method outperforms state of the art by (i) reducing the verified errors of trained models up to 95.64%; (ii) totally achieving up to 602.50x speedup; and (iii) scaling up to larger models with up to 138 million trainable parameters. The demo is available at https://github.com/zhangzhaodi233/ABSCERT.git.
****

Exploring Structured Semantic Prior for Multi Label Recognition With Incomplete Labels

Zixuan Ding, Ao Wang, Hui Chen, Qiang Zhang, Pengzhang Liu, Yongjun Bao, Weipeng Yan, Jungong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3398-3407

Multi-label recognition (MLR) with incomplete labels is very challenging. Recent works strive to explore the image-to-label correspondence in the vision-language model, i.e., CLIP, to compensate for insufficient annotations. In spite of promising performance, they generally overlook the valuable prior about the label-to-label correspondence. In this paper, we advocate remedying the deficiency of label supervision for the MLR with incomplete labels by deriving a structured semantic prior about the label-to-label correspondence via a semantic prior prompter. We then present a novel Semantic Correspondence Prompt Network (SCPNet), which can thoroughly explore the structured semantic prior. A Prior-Enhanced Self-Supervised Learning method is further introduced to enhance the use of the prior. Comprehensive experiments and analyses on several widely used benchmark datasets show that our method significantly outperforms existing methods on all datasets, well demonstrating the effectiveness and the superiority of our method. Our code will be available at https://github.com/jameslahm/SCPNet.
****

Instance-Specific and Model-Adaptive Supervision for Semi-Supervised Semantic Segmentation

Zhen Zhao, Sifan Long, Jimin Pi, Jingdong Wang, Luping Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23705-23714

Recently, semi-supervised semantic segmentation has achieved promising performance with a small fraction of labeled data. However, most existing studies treat all unlabeled data equally and barely consider the differences and training difficulties among unlabeled instances. Differentiating unlabeled instances can promote instance-specific supervision to adapt to the model's evolution dynamically. In this paper, we emphasize the cruciality of instance differences and propose an instance-specific and model-adaptive supervision for semi-supervised semantic segmentation, named iMAS. Relying on the model's performance, iMAS employs a class-weighted symmetric intersection-over-union to evaluate quantitative hardness of each unlabeled instance and supervises the training on unlabeled data in a model-adaptive manner. Specifically, iMAS learns from unlabeled instances progressively by weighing their corresponding consistency losses based on the evaluated hardness. Besides, iMAS dynamically adjusts the augmentation for each instance such that the distortion degree of augmented instances is adapted to the model's generalization capability across the training course. Not integrating additional losses and training procedures, iMAS can obtain remarkable performance gains against current state-of-the-art approaches on segmentation benchmarks under different semi-supervised partition protocols.
****

3D Shape Reconstruction of Semi-Transparent Worms

Thomas P. Ilett, Omer Yuval, Thomas Ranner, Netta Cohen, David C. Hogg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

, 2023, pp. 12565-12575
3D shape reconstruction typically requires identifying object features or textures in multiple images of a subject. This approach is not viable when the subject is semi-transparent and moving in and out of focus. Here we overcome these challenges by rendering a candidate shape with adaptive blurring and transparency for comparison with the images. We use the microscopic nematode Caenorhabditis elegans as a case study as it freely explores a 3D complex fluid with constantly changing optical properties. We model the slender worm as a 3D curve using an intrinsic parametrisation that naturally admits biologically-informed constraints and regularisation. To account for the changing optics we develop a novel differentiable renderer to construct images from 2D projections and compare against raw images to generate a pixel-wise error to jointly update the curve, camera and renderer parameters using gradient descent. The method is robust to interference such as bubbles and dirt trapped in the fluid, stays consistent through complex sequences of postures, recovers reliable estimates from blurry images and provides a significant improvement on previous attempts to track C. elegans in 3D. Our results demonstrate the potential of direct approaches to shape estimation in complex physical environments in the absence of ground-truth data.
****
Mapping Degeneration Meets Label Evolution: Learning Infrared Small Target Detection With Single Point Supervision
Xinyi Ying, Li Liu, Yingqian Wang, Ruojing Li, Nuo Chen, Zaiping Lin, Weidong Sheng, Shilin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15528-15538
Training a convolutional neural network (CNN) to detect infrared small targets in a fully supervised manner has gained remarkable research interests in recent years, but is highly labor expensive since a large number of per-pixel annotations are required. To handle this problem, in this paper, we make the first attempt to achieve infrared small target detection with point-level supervision. Interestingly, during the training phase supervised by point labels, we discover that CNNs first learn to segment a cluster of pixels near the targets, and then gradually converge to predict groundtruth point labels. Motivated by this "mapping degeneration" phenomenon, we propose a label evolution framework named label evolution with single point supervision (LESPS) to progressively expand the point label by leveraging the intermediate predictions of CNNs. In this way, the network predictions can finally approximate the updated pseudo labels, and a pixel-level target mask can be obtained to train CNNs in an end-to-end manner. We conduct extensive experiments with insightful visualizations to validate the effectiveness of our method. Experimental results show that CNNs equipped with LESPS can well recover the target masks from corresponding point labels, and can achieve over 70% and 95% of their fully supervised performance in terms of pixel-level intersection over union (IoU) and object-level probability of detection (Pd), respectively. Code is available at https://github.com/XinyiYing/LESPS.
****
Swept-Angle Synthetic Wavelength Interferometry
Alankar Kotwal, Anat Levin, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8233-8243
We present a new imaging technique, swept-angle synthetic wavelength interferometry, for full-field micron-scale 3D sensing. As in conventional synthetic wavelength interferometry, our technique uses light consisting of two narrowly-separated optical wavelengths, resulting in per-pixel interferometric measurements whose phase encodes scene depth. Our technique additionally uses a new type of light source that, by emulating spatially-incoherent illumination, makes interferometric measurements insensitive to aberrations and (sub)surface scattering, effects that corrupt phase measurements. The resulting technique combines the robustness to such corruptions of scanning interferometric setups, with the speed of full-field interferometric setups. Overall, our technique can recover full-frame depth at a lateral and axial resolution of 5 microns, at frame rates of 5 Hz, even under strong ambient light. We build an experimental prototype, and use it to demonstrate these capabilities by scanning a variety of objects, including objects

representative of applications in inspection and fabrication, and objects that contain challenging light scattering effects.
****

Delving Into Shape-Aware Zero-Shot Semantic Segmentation
Xinyu Liu, Beiwen Tian, Zhen Wang, Rui Wang, Kehua Sheng, Bo Zhang, Hao Zhao, Guyue Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2999-3009
Thanks to the impressive progress of large-scale vision-language pretraining, recent recognition models can classify arbitrary objects in a zero-shot and open-set manner, with a surprisingly high accuracy. However, translating this success to semantic segmentation is not trivial, because this dense prediction task requires not only accurate semantic understanding but also fine shape delineation and existing vision-language models are trained with image-level language descriptions. To bridge this gap, we pursue shape-aware zero-shot semantic segmentation in this study. Inspired by classical spectral methods in the image segmentation literature, we propose to leverage the eigen vectors of Laplacian matrices constructed with self-supervised pixel-wise features to promote shape-awareness. Despite that this simple and effective technique does not make use of the masks of seen classes at all, we demonstrate that it out-performs a state-of-the-art shape-aware formulation that aligns ground truth and predicted edges during training. We also delve into the performance gains achieved on different datasets using different backbones and draw several interesting and conclusive observations: the benefits of promoting shape-awareness highly relates to mask compactness and language embedding locality. Finally, our method sets new state-of-the-art performance for zero-shot semantic segmentation on both Pascal and COCO, with significant margins. Code and models will be accessed at https://github.com/Liuxinyv/SAZS.
****

Post-Training Quantization on Diffusion Models
Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, Yan Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1972-1981
Denoising diffusion (score-based) generative models have recently achieved significant accomplishments in generating realistic and diverse data. These approaches define a forward diffusion process for transforming data into noise and a backward denoising process for sampling data from noise. Unfortunately, the generation process of current denoising diffusion models is notoriously slow due to the lengthy iterative noise estimations, which rely on cumbersome neural networks. It prevents the diffusion models from being widely deployed, especially on edge devices. Previous works accelerate the generation process of diffusion model (DM) via finding shorter yet effective sampling trajectories. However, they overlook the cost of noise estimation with a heavy network in every iteration. In this work, we accelerate generation from the perspective of compressing the noise estimation network. Due to the difficulty of retraining DMs, we exclude mainstream training-aware compression paradigms and introduce post-training quantization (PTQ) into DM acceleration. However, the output distributions of noise estimation networks change with time-step, making previous PTQ methods fail in DMs since they are designed for single-time step scenarios. To devise a DM-specific PTQ method, we explore PTQ on DM in three aspects: quantized operations, calibration dataset, and calibration metric. We summarize and use several observations derived from all-inclusive investigations to formulate our method, which especially targets the unique multi-time-step structure of DMs. Experimentally, our method can directly quantize full-precision DMs into 8-bit models while maintaining or even improving their performance in a training-free manner. Importantly, our method can serve as a plug-and-play module on other fast-sampling methods, such as DDIM.
****

Adaptive Global Decay Process for Event Cameras
Urbano Miguel Nunes, Ryad Benosman, Sio-Hoi Ieng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9771-9780
In virtually all event-based vision problems, there is the need to select the mo

st recent events, which are assumed to carry the most relevant information conte
nt. To achieve this, at least one of three main strategies is applied, namely: 1
) constant temporal decay or fixed time window, 2) constant number of events, an
d 3) flow-based lifetime of events. However, these strategies suffer from at lea
st one major limitation each. We instead propose a novel decay process for event
 cameras that adapts to the global scene dynamics and whose latency is in the or
der of nanoseconds. The main idea is to construct an adaptive quantity that enco
des the global scene dynamics, denoted by event activity. The proposed method is
 evaluated in several event-based vision problems and datasets, consistently imp
roving the corresponding baseline methods' performance. We thus believe it can h
ave a significant widespread impact on event-based research. Code available: htt
ps://github.com/neuromorphic-paris/event_batch.
****
Multi-Space Neural Radiance Fields
Ze-Xin Yin, Jiaxiong Qiu, Ming-Ming Cheng, Bo Ren; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12407-124
16
Neural Radiance Fields (NeRF) and its variants have reached state-of-the-art per
formance in many novel-view-synthesis-related tasks. However, current NeRF-based
 methods still suffer from the existence of reflective objects, often resulting
in blurry or distorted rendering. Instead of calculating a single radiance field
, we propose a multispace neural radiance field (MS-NeRF) that represents the sc
ene using a group of feature fields in parallel sub-spaces, which leads to a bet
ter understanding of the neural network toward the existence of reflective and r
efractive objects. Our multi-space scheme works as an enhancement to existing Ne
RF methods, with only small computational overheads needed for training and infe
rring the extra-space outputs. We demonstrate the superiority and compatibility
of our approach using three representative NeRF-based models, i.e., NeRF, Mip-Ne
RF, and Mip-NeRF 360. Comparisons are performed on a novelly constructed dataset
 consisting of 25 synthetic scenes and 7 real captured scenes with complex refle
ction and refraction, all having 360-degree viewpoints. Extensive experiments sh
ow that our approach significantly outperforms the existing single-space NeRF me
thods for rendering high-quality scenes concerned with complex light paths throu
gh mirror-like objects.
****
Leveraging Inter-Rater Agreement for Classification in the Presence of Noisy Lab
els
Maria Sofia Bucarelli, Lucas Cassano, Federico Siciliano, Amin Mantrach, Fabrizi
o Silvestri; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2023, pp. 3439-3448
In practical settings, classification datasets are obtained through a labelling
process that is usually done by humans. Labels can be noisy as they are obtained
 by aggregating the different individual labels assigned to the same sample by m
ultiple, and possibly disagreeing, annotators. The inter-rater agreement on thes
e datasets can be measured while the underlying noise distribution to which the
labels are subject is assumed to be unknown. In this work, we: (i) show how to l
everage the inter-annotator statistics to estimate the noise distribution to whi
ch labels are subject; (ii) introduce methods that use the estimate of the noise
 distribution to learn from the noisy dataset; and (iii) establish generalizatio
n bounds in the empirical risk minimization framework that depend on the estimat
ed quantities. We conclude the paper by providing experiments that illustrate ou
r findings.
****
Bitstream-Corrupted JPEG Images Are Restorable: Two-Stage Compensation and Align
ment Framework for Image Restoration
Wenyang Liu, Yi Wang, Kim-Hui Yap, Lap-Pui Chau; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9979-9988
In this paper, we study a real-world JPEG image restoration problem with bit err
ors on the encrypted bitstream. The bit errors bring unpredictable color casts a
nd block shifts on decoded image contents, which cannot be trivially resolved by

existing image restoration methods mainly relying on pre-defined degradation models in the pixel domain. To address these challenges, we propose a robust JPEG decoder, followed by a two-stage compensation and alignment framework to restore bitstream-corrupted JPEG images. Specifically, the robust JPEG decoder adopts an error-resilient mechanism to decode the corrupted JPEG bitstream. The two-stage framework is composed of the self-compensation and alignment (SCA) stage and the guided-compensation and alignment (GCA) stage. The SCA adaptively performs block-wise image color compensation and alignment based on the estimated color and block offsets via image content similarity. The GCA leverages the extracted low-resolution thumbnail from the JPEG header to guide full-resolution pixel-wise image restoration in a coarse-to-fine manner. It is achieved by a coarse-guided pix2pix network and a refine-guided bi-directional Laplacian pyramid fusion network. We conduct experiments on three benchmarks with varying degrees of bit error rates. Experimental results and ablation studies demonstrate the superiority of our proposed method. The code will be released at https://github.com/wenyang001/Two-ACIR.
****
Analyzing Physical Impacts Using Transient Surface Wave Imaging
Tianyuan Zhang, Mark Sheinin, Dorian Chan, Mark Rau, Matthew O’Toole, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4339-4348
The subtle vibrations on an object's surface contain information about the object's physical properties and its interaction with the environment. Prior works imaged surface vibration to recover the object's material properties via modal analysis, which discards the transient vibrations propagating immediately after the object is disturbed. Conversely, prior works that captured transient vibrations focused on recovering localized signals (e.g., recording nearby sound sources), neglecting the spatiotemporal relationship between vibrations at different object points. In this paper, we extract information from the transient surface vibrations simultaneously measured at a sparse set of object points using the dual-shutter camera described by Sheinin[31]. We model the geometry of an elastic wave generated shortly after an object's surface is disturbed (e.g., a knock or a footstep), and use the model to localize the disturbance source for various materials (e.g., wood, plastic, tile). We also show that transient object vibrations contain additional cues about the impact force and the impacting object's material properties. We demonstrate our approach in applications like localizing the strikes of a ping-pong ball on a table mid-play and recovering the footsteps' locations by imaging the floor vibrations they create.
****
X-Pruner: eXplainable Pruning for Vision Transformers
Lu Yu, Wei Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24355-24363
Recently vision transformer models have become prominent models for a range of tasks. These models, however, usually suffer from intensive computational costs and heavy memory requirements, making them impractical for deployment on edge platforms. Recent studies have proposed to prune transformers in an unexplainable manner, which overlook the relationship between internal units of the model and the target class, thereby leading to inferior performance. To alleviate this problem, we propose a novel explainable pruning framework dubbed X-Pruner, which is designed by considering the explainability of the pruning criterion. Specifically, to measure each prunable unit's contribution to predicting each target class, a novel explainability-aware mask is proposed and learned in an end-to-end manner. Then, to preserve the most informative units and learn the layer-wise pruning rate, we adaptively search the layer-wise threshold that differentiates between unpruned and pruned units based on their explainability-aware mask values. To verify and evaluate our method, we apply the X-Pruner on representative transformer models including the DeiT and Swin Transformer. Comprehensive simulation results demonstrate that the proposed X-Pruner outperforms the state-of-the-art black-box methods with significantly reduced computational costs and slight performance degradation.

****
Hard Sample Matters a Lot in Zero-Shot Quantization
Huantong Li, Xiangmiao Wu, Fanbing Lv, Daihai Liao, Thomas H. Li, Yonggang Zhang, Bo Han, Mingkui Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24417-24426
Zero-shot quantization (ZSQ) is promising for compressing and accelerating deep neural networks when the data for training full-precision models are inaccessible. In ZSQ, network quantization is performed using synthetic samples, thus, the performance of quantized models depends heavily on the quality of synthetic samples. Nonetheless, we find that the synthetic samples constructed in existing ZSQ methods can be easily fitted by models. Accordingly, quantized models obtained by these methods suffer from significant performance degradation on hard samples. To address this issue, we propose HArd sample Synthesizing and Training (HAST). Specifically, HAST pays more attention to hard samples when synthesizing samples and makes synthetic samples hard to fit when training quantized models. HAST aligns features extracted by full-precision and quantized models to ensure the similarity between features extracted by these two models. Extensive experiments show that HAST significantly outperforms existing ZSQ methods, achieving performance comparable to models that are quantized with real data.
****
Meta Compositional Referring Expression Segmentation
Li Xu, Mark He Huang, Xindi Shang, Zehuan Yuan, Ying Sun, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19478-19487
Referring expression segmentation aims to segment an object described by a language expression from an image. Despite the recent progress on this task, existing models tackling this task may not be able to fully capture semantics and visual representations of individual concepts, which limits their generalization capability, especially when handling novel compositions of learned concepts. In this work, through the lens of meta learning, we propose a Meta Compositional Referring Expression Segmentation (MCRES) framework to enhance model compositional generalization performance. Specifically, to handle various levels of novel compositions, our framework first uses training data to construct a virtual training set and multiple virtual testing sets, where data samples in each virtual testing set contain a level of novel compositions w.r.t. the support set. Then, following a novel meta optimization scheme to optimize the model to obtain good testing performance on the virtual testing sets after training on the virtual training set, our framework can effectively drive the model to better capture semantics and visual representations of individual concepts, and thus obtain robust generalization performance even when handling novel compositions. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our framework.
****
Histopathology Whole Slide Image Analysis With Heterogeneous Graph Representation Learning
Tsai Hor Chan, Fernando Julio Cendra, Lan Ma, Guosheng Yin, Lequan Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15661-15670
Graph-based methods have been extensively applied to whole slide histopathology image (WSI) analysis due to the advantage of modeling the spatial relationships among different entities. However, most of the existing methods focus on modeling WSIs with homogeneous graphs (e.g., with homogeneous node type). Despite their successes, these works are incapable of mining the complex structural relations between biological entities (e.g., the diverse interaction among different cell types) in the WSI. We propose a novel heterogeneous graph-based framework to leverage the inter-relationships among different types of nuclei for WSI analysis. Specifically, we formulate the WSI as a heterogeneous graph with "nucleus-type" attribute to each node and a semantic similarity attribute to each edge. We then present a new heterogeneous-graph edge attribute transformer (HEAT) to take advantage of the edge and node heterogeneity during massage aggregating. Further, we design a new pseudo-label-based semantic-consistent pooling mechanism to obta

in graph-level features, which can mitigate the over-parameterization issue of c
onventional cluster-based pooling. Additionally, observing the limitations of ex
isting association-based localization methods, we propose a causal-driven approa
ch attributing the contribution of each node to improve the interpretability of
our framework. Extensive experiments on three public TCGA benchmark datasets dem
onstrate that our framework outperforms the state-of-the-art methods with consid
erable margins on various tasks. Our codes are available at https://github.com/H
KU-MedAI/WSI-HGNN.
****
ScanDMM: A Deep Markov Model of Scanpath Prediction for 360deg Images
Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, Zhou Wang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp
. 6989-6999
Scanpath prediction for 360deg images aims to produce dynamic gaze behaviors bas
ed on the human visual perception mechanism. Most existing scanpath prediction m
ethods for 360deg images do not give a complete treatment of the time-dependency
 when predicting human scanpath, resulting in inferior performance and poor gene
ralizability. In this paper, we present a scanpath prediction method for 360deg
images by designing a novel Deep Markov Model (DMM) architecture, namely ScanDMM
. We propose a semantics-guided transition function to learn the nonlinear dynam
ics of time-dependent attentional landscape. Moreover, a state initialization st
rategy is proposed by considering the starting point of viewing, enabling the mo
del to learn the dynamics with the correct "launcher". We further demonstrate th
at our model achieves state-of-the-art performance on four 360deg image database
s, and exhibit its generalizability by presenting two applications of applying s
canpath prediction models to other visual tasks - saliency detection and image q
uality assessment, expecting to provide profound insights into these fields.
****
Towards All-in-One Pre-Training via Maximizing Multi-Modal Mutual Information
Weijie Su, Xizhou Zhu, Chenxin Tao, Lewei Lu, Bin Li, Gao Huang, Yu Qiao, Xiaoga
ng Wang, Jie Zhou, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2023, pp. 15888-15899
To effectively exploit the potential of large-scale models, various pre-training
 strategies supported by massive data from different sources are proposed, inclu
ding supervised pre-training, weakly-supervised pre-training, and self-supervise
d pre-training. It has been proved that combining multiple pre-training strategi
es and data from various modalities/sources can greatly boost the training of la
rge-scale models. However, current works adopt a multi-stage pre-training system
, where the complex pipeline may increase the uncertainty and instability of the
 pre-training. It is thus desirable that these strategies can be integrated in a
 single-stage manner. In this paper, we first propose a general multi-modal mutu
al information formula as a unified optimization target and demonstrate that all
 mainstream approaches are special cases of our framework. Under this unified pe
rspective, we propose an all-in-one single-stage pre-training approach, named Ma
ximizing Multi-modal Mutual Information Pre-training (M3I Pre-training). Our app
roach achieves better performance than previous pre-training methods on various
vision benchmarks, including ImageNet classification, COCO object detection, LVI
S long-tailed object detection, and ADE20k semantic segmentation. Notably, we su
ccessfully pre-train a billion-level parameter image backbone and achieve state-
of-the-art performance on various benchmarks under public data setting. Code sha
ll be released at https://github.com/OpenGVLab/M3I-Pretraining.
****
Aligning Bag of Regions for Open-Vocabulary Object Detection
Size Wu, Wenwei Zhang, Sheng Jin, Wentao Liu, Chen Change Loy; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 15254-15264
Pre-trained vision-language models (VLMs) learn to align vision and language rep
resentations on large-scale datasets, where each image-text pair usually contain
s a bag of semantic concepts. However, existing open-vocabulary object detectors
 only align region embeddings individually with the corresponding features extra

cted from the VLMs. Such a design leaves the compositional structure of semantic concepts in a scene under-exploited, although the structure may be implicitly learned by the VLMs. In this work, we propose to align the embedding of bag of regions beyond individual regions. The proposed method groups contextually interrelated regions as a bag. The embeddings of regions in a bag are treated as embeddings of words in a sentence, and they are sent to the text encoder of a VLM to obtain the bag-of-regions embedding, which is learned to be aligned to the corresponding features extracted by a frozen VLM. Applied to the commonly used Faster R-CNN, our approach surpasses the previous best results by 4.6 box AP 50 and 2.8 mask AP on novel categories of open-vocabulary COCO and LVIS benchmarks, respectively. Code and models are available at https://github.com/wusize/ovdet.
****
Two-View Geometry Scoring Without Correspondences
Axel Barroso-Laguna, Eric Brachmann, Victor Adrian Prisacariu, Gabriel J. Brostow, Daniyar Turmukhambetov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8979-8989
Camera pose estimation for two-view geometry traditionally relies on RANSAC. Normally, a multitude of image correspondences leads to a pool of proposed hypotheses, which are then scored to find a winning model. The inlier count is generally regarded as a reliable indicator of "consensus". We examine this scoring heuristic, and find that it favors disappointing models under certain circumstances. As a remedy, we propose the Fundamental Scoring Network (FSNet), which infers a score for a pair of overlapping images and any proposed fundamental matrix. It does not rely on sparse correspondences, but rather embodies a two-view geometry model through an epipolar attention mechanism that predicts the pose error of the two images. FSNet can be incorporated into traditional RANSAC loops. We evaluate FSNet on fundamental and essential matrix estimation on indoor and outdoor datasets, and establish that FSNet can successfully identify good poses for pairs of images with few or unreliable correspondences. Besides, we show that naively combining FSNet with MAGSAC++ scoring approach achieves state of the art results.
****
Annealing-Based Label-Transfer Learning for Open World Object Detection
Yuqing Ma, Hainan Li, Zhange Zhang, Jinyang Guo, Shanghang Zhang, Ruihao Gong, Xianglong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11454-11463
Open world object detection (OWOD) has attracted extensive attention due to its practicability in the real world. Previous OWOD works manually designed unknown-discover strategies to select unknown proposals from the background, suffering from uncertainties without appropriate priors. In this paper, we claim the learning of object detection could be seen as an object-level feature-entanglement process, where unknown traits are propagated to the known proposals through convolutional operations and could be distilled to benefit unknown recognition without manual selection. Therefore, we propose a simple yet effective Annealing-based Label-Transfer framework, which sufficiently explores the known proposals to alleviate the uncertainties. Specifically, a Label-Transfer Learning paradigm is introduced to decouple the known and unknown features, while a Sawtooth Annealing Scheduling strategy is further employed to rebuild the decision boundaries of the known and unknown classes, thus promoting both known and unknown recognition. Moreover, previous OWOD works neglected the trade-off of known and unknown performance, and we thus introduce a metric called Equilibrium Index to comprehensively evaluate the effectiveness of the OWOD models. To the best of our knowledge, this is the first OWOD work without manual unknown selection. Extensive experiments conducted on the common-used benchmark validate that our model achieves superior detection performance (200% unknown mAP improvement with the even higher known detection performance) compared to other state-of-the-art methods. Our code is available at https://github.com/DIG-Beihang/ALLOW.git.
****
Continual Semantic Segmentation With Automatic Memory Sample Selection
Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.

3082-3092
Continual Semantic Segmentation (CSS) extends static semantic segmentation by in
crementally introducing new classes for training. To alleviate the catastrophic
forgetting issue in CSS, a memory buffer that stores a small number of samples f
rom the previous classes is constructed for replay. However, existing methods se
lect the memory samples either randomly or based on a single-factor-driven hand-
crafted strategy, which has no guarantee to be optimal. In this work, we propose
a novel memory sample selection mechanism that selects informative samples for
effective replay in a fully automatic way by considering comprehensive factors i
ncluding sample diversity and class performance. Our mechanism regards the selec
tion operation as a decision-making process and learns an optimal selection poli
cy that directly maximizes the validation performance on a reward set. To facili
tate the selection decision, we design a novel state representation and a dual-s
tage action space. Our extensive experiments on Pascal-VOC 2012 and ADE 20K data
sets demonstrate the effectiveness of our approach with state-of-the-art (SOTA)
performance achieved, outperforming the second-place one by 12.54% for the 6-sta
ge setting on Pascal-VOC 2012.
****
Meta-Tuning Loss Functions and Data Augmentation for Few-Shot Object Detection
Berkan Demirel, Orhun Bu■ra Baran, Ramazan Gokberk Cinbis; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7
339-7349
Few-shot object detection, the problem of modelling novel object detection categ
ories with few training instances, is an emerging topic in the area of few-shot
learning and object detection. Contemporary techniques can be divided into two g
roups: fine-tuning based and meta-learning based approaches. While meta-learning
approaches aim to learn dedicated meta-models for mapping samples to novel clas
s models, fine-tuning approaches tackle few-shot detection in a simpler manner,
by adapting the detection model to novel classes through gradient based optimiza
tion. Despite their simplicity, fine-tuning based approaches typically yield com
petitive detection results. Based on this observation, we focus on the role of l
oss functions and augmentations as the force driving the fine-tuning process, an
d propose to tune their dynamics through meta-learning principles. The proposed
training scheme, therefore, allows learning inductive biases that can boost few-
shot detection, while keeping the advantages of fine-tuning based approaches. In
addition, the proposed approach yields interpretable loss functions, as opposed
to highly parametric and complex few-shot meta-models. The experimental results
highlight the merits of the proposed scheme, with significant improvements over
the strong fine-tuning based few-shot detection baselines on benchmark Pascal V
OC and MS-COCO datasets, in terms of both standard and generalized few-shot perf
ormance metrics.
****
A Light Weight Model for Active Speaker Detection
Junhua Liao, Haihan Duan, Kanghui Feng, Wanbing Zhao, Yanbing Yang, Liangyin Che
n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2023, pp. 22932-22941
Active speaker detection is a challenging task in audio-visual scenarios, with t
he aim to detect who is speaking in one or more speaker scenarios. This task has
received considerable attention because it is crucial in many applications. Exi
sting studies have attempted to improve the performance by inputting multiple ca
ndidate information and designing complex models. Although these methods have ac
hieved excellent performance, their high memory and computational power consumpt
ion render their application to resource-limited scenarios difficult. Therefore,
in this study, a lightweight active speaker detection architecture is construct
ed by reducing the number of input candidates, splitting 2D and 3D convolutions
for audio-visual feature extraction, and applying gated recurrent units with low
computational complexity for cross-modal modeling. Experimental results on the
AVA-ActiveSpeaker dataset reveal that the proposed framework achieves competitiv
e mAP performance (94.1% vs. 94.2%), while the resource costs are significantly
lower than the state-of-the-art method, particularly in model parameters (1.0M v

s. 22.5M, approximately 23x) and FLOPs (0.6G vs. 2.6G, approximately 4x). Additionally, the proposed framework also performs well on the Columbia dataset, thus demonstrating good robustness. The code and model weights are available at https://github.com/Junhua-Liao/Light-ASD.
****
Self-Supervised Video Forensics by Audio-Visual Anomaly Detection
Chao Feng, Ziyang Chen, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10491-10503
Manipulated videos often contain subtle inconsistencies between their visual and audio signals. We propose a video forensics method, based on anomaly detection, that can identify these inconsistencies, and that can be trained solely using real, unlabeled data. We train an autoregressive model to generate sequences of audio-visual features, using feature sets that capture the temporal synchronization between video frames and sound. At test time, we then flag videos that the model assigns low probability. Despite being trained entirely on real videos, our model obtains strong performance on the task of detecting manipulated speech videos. Project site: https://cfeng16.github.io/audio-visual-forensics.
****
CLIP2Scene: Towards Label-Efficient 3D Scene Understanding by CLIP
Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7020-7030
Contrastive Language-Image Pre-training (CLIP) achieves promising results in 2D zero-shot and few-shot learning. Despite the impressive performance in 2D, applying CLIP to help the learning in 3D scene understanding has yet to be explored. In this paper, we make the first attempt to investigate how CLIP knowledge benefits 3D scene understanding. We propose CLIP2Scene, a simple yet effective framework that transfers CLIP knowledge from 2D image-text pre-trained models to a 3D point cloud network. We show that the pre-trained 3D network yields impressive performance on various downstream tasks, i.e., annotation-free and fine-tuning with labelled data for semantic segmentation. Specifically, built upon CLIP, we design a Semantic-driven Cross-modal Contrastive Learning framework that pre-trains a 3D network via semantic and spatial-temporal consistency regularization. For the former, we first leverage CLIP's text semantics to select the positive and negative point samples and then employ the contrastive loss to train the 3D network. In terms of the latter, we force the consistency between the temporally coherent point cloud features and their corresponding image features. We conduct experiments on SemanticKITTI, nuScenes, and ScanNet. For the first time, our pre-trained network achieves annotation-free 3D semantic segmentation with 20.8% and 25.08% mIoU on nuScenes and ScanNet, respectively. When fine-tuned with 1% or 100% labelled data, our method significantly outperforms other self-supervised methods, with improvements of 8% and 1% mIoU, respectively. Furthermore, we demonstrate the generalizability for handling cross-domain datasets. Code is publicly available.
****
GCFAgg: Global and Cross-View Feature Aggregation for Multi-View Clustering
Weiqing Yan, Yuanyang Zhang, Chenlei Lv, Chang Tang, Guanghui Yue, Liang Liao, Weisi Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19863-19872
Multi-view clustering can partition data samples into their categories by learning a consensus representation in unsupervised way and has received more and more attention in recent years. However, most existing deep clustering methods learn consensus representation or view-specific representations from multiple views via view-wise aggregation way, where they ignore structure relationship of all samples. In this paper, we propose a novel multi-view clustering network to address these problems, called Global and Cross-view Feature Aggregation for Multi-View Clustering (GCFAggMVC). Specifically, the consensus data presentation from multiple views is obtained via cross-sample and cross-view feature aggregation, which fully explores the complementary of similar samples. Moreover, we align the consensus representation and the view-specific representation by the structure-gu

ided contrastive learning module, which makes the view-specific representations from different samples with high structure relationship similar. The proposed module is a flexible multi-view data representation module, which can be also embedded to the incomplete multi-view data clustering task via plugging our module into other frameworks. Extensive experiments show that the proposed method achieves excellent performance in both complete multi-view data clustering tasks and incomplete multi-view data clustering tasks.
****
Class Balanced Adaptive Pseudo Labeling for Federated Semi-Supervised Learning
Ming Li, Qingli Li, Yan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16292-16301
This paper focuses on federated semi-supervised learning (FSSL), assuming that few clients have fully labeled data (labeled clients) and the training datasets in other clients are fully unlabeled (unlabeled clients). Existing methods attempt to deal with the challenges caused by not independent and identically distributed data (Non-IID) setting. Though methods such as sub-consensus models have been proposed, they usually adopt standard pseudo labeling or consistency regularization on unlabeled clients which can be easily influenced by imbalanced class distribution. Thus, problems in FSSL are still yet to be solved. To seek for a fundamental solution to this problem, we present Class Balanced Adaptive Pseudo Labeling (CBAFed), to study FSSL from the perspective of pseudo labeling. In CBAFed, the first key element is a fixed pseudo labeling strategy to handle the catastrophic forgetting problem, where we keep a fixed set by letting pass information of unlabeled data at the beginning of the unlabeled client training in each communication round. The second key element is that we design class balanced adaptive thresholds via considering the empirical distribution of all training data in local clients, to encourage a balanced training process. To make the model reach a better optimum, we further propose a residual weight connection in local supervised training and global model aggregation. Extensive experiments on five datasets demonstrate the superiority of CBAFed. Code will be released.
****
Rethinking Out-of-Distribution (OOD) Detection: Masked Image Modeling Is All You Need
Jingyao Li, Pengguang Chen, Zexin He, Shaozuo Yu, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11578-11589
The core of out-of-distribution (OOD) detection is to learn the in-distribution (ID) representation, which is distinguishable from OOD samples. Previous work applied recognition-based methods to learn the ID features, which tend to learn shortcuts instead of comprehensive representations. In this work, we find surprisingly that simply using reconstruction-based methods could boost the performance of OOD detection significantly. We deeply explore the main contributors of OOD detection and find that reconstruction-based pretext tasks have the potential to provide a generally applicable and efficacious prior, which benefits the model in learning intrinsic data distributions of the ID dataset. Specifically, we take Masked Image Modeling as a pretext task for our OOD detection framework (MOOD). Without bells and whistles, MOOD outperforms previous SOTA of one-class OOD detection by 5.7%, multi-class OOD detection by 3.0%, and near-distribution OOD detection by 2.1%. It even defeats the 10-shot-per-class outlier exposure OOD detection, although we do not include any OOD samples for our detection.
****
DeGPR: Deep Guided Posterior Regularization for Multi-Class Cell Detection and Counting
Aayush Kumar Tyagi, Chirag Mohapatra, Prasenjit Das, Govind Makharia, Lalita Mehra, Prathosh AP, Mausam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23913-23923
Multi-class cell detection and counting is an essential task for many pathological diagnoses. Manual counting is tedious and often leads to inter-observer variations among pathologists. While there exist multiple, general-purpose, deep learning-based object detection and counting methods, they may not readily transfer

to detecting and counting cells in medical images, due to the limited data, presence of tiny overlapping objects, multiple cell types, severe class-imbalance, minute differences in size/shape of cells, etc. In response, we propose guided posterior regularization DeGPR, which assists an object detector by guiding it to exploit discriminative features among cells. The features may be pathologist-provided or inferred directly from visual data. We validate our model on two publicly available datasets (CoNSeP and MoNuSAC), and on MuCeD, a novel dataset that we contribute. MuCeD consists of 55 biopsy images of the human duodenum for predicting celiac disease. We perform extensive experimentation with three object detection baselines on three datasets to show that DeGPR is model-agnostic, and consistently improves baselines obtaining up to 9% (absolute) mAP gains.
****
Masked Scene Contrast: A Scalable Framework for Unsupervised 3D Representation Learning
Xiaoyang Wu, Xin Wen, Xihui Liu, Hengshuang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9415-9424
As a pioneering work, PointContrast conducts unsupervised 3D representation learning via leveraging contrastive learning over raw RGB-D frames and proves its effectiveness on various downstream tasks. However, the trend of large-scale unsupervised learning in 3D has yet to emerge due to two stumbling blocks: the inefficiency of matching RGB-D frames as contrastive views and the annoying mode collapse phenomenon mentioned in previous works. Turning the two stumbling blocks into empirical stepping stones, we first propose an efficient and effective contrastive learning framework, which generates contrastive views directly on scene-level point clouds by a well-curated data augmentation pipeline and a practical view mixing strategy. Second, we introduce reconstructive learning on the contrastive learning framework with an exquisite design of contrastive cross masks, which targets the reconstruction of point color and surfel normal. Our Masked Scene Contrast (MSC) framework is capable of extracting comprehensive 3D representations more efficiently and effectively. It accelerates the pre-training procedure by at least 3x and still achieves an uncompromised performance compared with previous work. Besides, MSC also enables large-scale 3D pre-training across multiple datasets, which further boosts the performance and achieves state-of-the-art fine-tuning results on several downstream tasks, e.g., 75.5% mIoU on ScanNet semantic segmentation validation set.
****
Multi Domain Learning for Motion Magnification
Jasdeep Singh, Subrahmanyam Murala, G. Sankara Raju Kosuru; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13914-13923
Video motion magnification makes subtle invisible motions visible, such as small chest movements while breathing, subtle vibrations in the moving objects etc. But small motions are prone to noise, illumination changes, large motions, etc. making the task difficult. Most state-of-the-art methods use hand-crafted concepts which result in small magnification, ringing artifacts etc. The deep learning based approach has higher magnification but is prone to severe artifacts in some scenarios. We propose a new phase based deep network for video motion magnification that operates in both domains (frequency and spatial) to address this issue. It generates motion magnification from frequency domain phase fluctuations and then improves its quality in the spatial domain. The proposed models are lightweight networks with fewer parameters (0.11M and   0.05M). Further, the proposed networks performance is compared to the SOTA approaches and evaluated on real-world and synthetic videos. Finally, an ablation study is also conducted to show the impact of different parts of the network.
****
LOGO: A Long-Form Video Dataset for Group Action Quality Assessment
Shiyi Zhang, Wenxun Dai, Sujia Wang, Xiangwei Shen, Jiwen Lu, Jie Zhou, Yansong Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2405-2414
Action quality assessment (AQA) has become an emerging topic since it can be ext

ensively applied in numerous scenarios. However, most existing methods and datasets focus on single-person short-sequence scenes, hindering the application of AQA in more complex situations. To address this issue, we construct a new multi-person long-form video dataset for action quality assessment named LOGO. Distinguished in scenario complexity, our dataset contains 200 videos from 26 artistic swimming events with 8 athletes in each sample along with an average duration of 204.2 seconds. As for richness in annotations, LOGO includes formation labels to depict group information of multiple athletes and detailed annotations on action procedures. Furthermore, we propose a simple yet effective method to model relations among athletes and reason about the potential temporal logic in long-form videos. Specifically, we design a group-aware attention module, which can be easily plugged into existing AQA methods, to enrich the clip-wise representations based on contextual group information. To benchmark LOGO, we systematically conduct investigations on the performance of several popular methods in AQA and action segmentation. The results reveal the challenges our dataset brings. Extensive experiments also show that our approach achieves state-of-the-art on the LOGO dataset. The dataset and code will be released at https://github.com/shiyi-zh0408/LOGO.
****
A Simple Baseline for Video Restoration With Grouped Spatial-Temporal Shift
Dasong Li, Xiaoyu Shi, Yi Zhang, Ka Chun Cheung, Simon See, Xiaogang Wang, Hongwei Qin, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9822-9832
Video restoration, which aims to restore clear frames from degraded videos, has numerous important applications. The key to video restoration depends on utilizing inter-frame information. However, existing deep learning methods often rely on complicated network architectures, such as optical flow estimation, deformable convolution, and cross-frame self-attention layers, resulting in high computational costs. In this study, we propose a simple yet effective framework for video restoration. Our approach is based on grouped spatial-temporal shift, which is a lightweight and straightforward technique that can implicitly capture inter-frame correspondences for multi-frame aggregation. By introducing grouped spatial shift, we attain expansive effective receptive fields. Combined with basic 2D convolution, this simple framework can effectively aggregate inter-frame information. Extensive experiments demonstrate that our framework outperforms the previous state-of-the-art method, while using less than a quarter of its computational cost, on both video deblurring and video denoising tasks. These results indicate the potential for our approach to significantly reduce computational overhead while maintaining high-quality results. Code is avaliable at https://github.com/dasongli1/Shift-Net.
****
UniSim: A Neural Closed-Loop Sensor Simulator
Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1389-1399
Rigorously testing autonomy systems is essential for making safe self-driving vehicles (SDV) a reality. It requires one to generate safety critical scenarios beyond what can be collected safely in the world, as many scenarios happen rarely on our roads. To accurately evaluate performance, we need to test the SDV on these scenarios in closed-loop, where the SDV and other actors interact with each other at each timestep. Previously recorded driving logs provide a rich resource to build these new scenarios from, but for closed loop evaluation, we need to modify the sensor data based on the new scene configuration and the SDV's decisions, as actors might be added or removed and the trajectories of existing actors and the SDV will differ from the original log. In this paper, we present UniSim, a neural sensor simulator that takes a single recorded log captured by a sensor-equipped vehicle and converts it into a realistic closed-loop multi-sensor simulation. UniSim builds neural feature grids to reconstruct both the static background and dynamic actors in the scene, and composites them together to simulate LiDAR and camera data at new viewpoints, with actors added or removed and at new p

lacements. To better handle extrapolated views, we incorporate learnable priors for dynamic objects, and leverage a convolutional network to complete unseen regions. Our experiments show UniSim can simulate realistic sensor data with small domain gap on downstream tasks. With UniSim, we demonstrate, for the first time, closed-loop evaluation of an autonomy system on safety-critical scenarios as if it were in the real world.
****
itKD: Interchange Transfer-Based Knowledge Distillation for 3D Object Detection
Hyeon Cho, Junyong Choi, Geonwoo Baek, Wonjun Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13540-13549
Point-cloud based 3D object detectors recently have achieved remarkable progress. However, most studies are limited to the development of network architectures for improving only their accuracy without consideration of the computational efficiency. In this paper, we first propose an autoencoder-style framework comprising channel-wise compression and decompression via interchange transfer-based knowledge distillation. To learn the map-view feature of a teacher network, the features from teacher and student networks are independently passed through the shared autoencoder; here, we use a compressed representation loss that binds the channel-wised compression knowledge from both student and teacher networks as a kind of regularization. The decompressed features are transferred in opposite directions to reduce the gap in the interchange reconstructions. Lastly, we present an head attention loss to match the 3D object detection information drawn by the multi-head self-attention mechanism. Through extensive experiments, we verify that our method can train the lightweight model that is well-aligned with the 3D point cloud detection task and we demonstrate its superiority using the well-known public datasets; e.g., Waymo and nuScenes.
****
SliceMatch: Geometry-Guided Aggregation for Cross-View Pose Estimation
Ted Lentsch, Zimin Xia, Holger Caesar, Julian F. P. Kooij; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17225-17234
This work addresses cross-view camera pose estimation, i.e., determining the 3-Degrees-of-Freedom camera pose of a given ground-level image w.r.t. an aerial image of the local area. We propose SliceMatch, which consists of ground and aerial feature extractors, feature aggregators, and a pose predictor. The feature extractors extract dense features from the ground and aerial images. Given a set of candidate camera poses, the feature aggregators construct a single ground descriptor and a set of pose-dependent aerial descriptors. Notably, our novel aerial feature aggregator has a cross-view attention module for ground-view guided aerial feature selection and utilizes the geometric projection of the ground camera's viewing frustum on the aerial image to pool features. The efficient construction of aerial descriptors is achieved using precomputed masks. SliceMatch is trained using contrastive learning and pose estimation is formulated as a similarity comparison between the ground descriptor and the aerial descriptors. Compared to the state-of-the-art, SliceMatch achieves a 19% lower median localization error on the VIGOR benchmark using the same VGG16 backbone at 150 frames per second, and a 50% lower error when using a ResNet50 backbone.
****
2PCNet: Two-Phase Consistency Training for Day-to-Night Unsupervised 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), 2023, pp. 11484-11493
Object detection at night is a challenging problem due to the absence of night image annotations. Despite several domain adaptation methods, achieving high-precision results remains an issue. False-positive error propagation is still observed in methods using the well-established student-teacher framework, particularly for small-scale and low-light objects. This paper proposes a two-phase consistency unsupervised domain adaptation network, 2PCNet, to address these issues. The

network employs high-confidence bounding-box predictions from the teacher in the first phase and appends them to the student's region proposals for the teacher to re-evaluate in the second phase, resulting in a combination of high and low confidence pseudo-labels. The night images and pseudo-labels are scaled-down before being used as input to the student, providing stronger small-scale pseudo-labels. To address errors that arise from low-light regions and other night-related attributes in images, we propose a night-specific augmentation pipeline called NightAug. This pipeline involves applying random augmentations, such as glare, blur, and noise, to daytime images. Experiments on publicly available datasets demonstrate that our method achieves superior results to state-of-the-art methods by 20%, and to supervised models trained directly on the target data.
****

Prefix Conditioning Unifies Language and Label Supervision

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, Tomas Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2861-2870

Pretraining visual models on web-scale image-caption datasets has recently emerged as a powerful alternative to traditional pretraining on image classification data. Image-caption datasets are more "open-domain", containing broader scene types and vocabulary words, and result in models that have strong performance in few- and zero-shot recognition tasks. However large-scale classification datasets can provide fine-grained categories with a balanced label distribution. In this work, we study a pretraining strategy that uses both classification and caption datasets to unite their complementary benefits. First, we show that naively unifying the datasets results in sub-optimal performance in downstream zero-shot recognition tasks, as the model is affected by dataset bias: the coverage of image domains and vocabulary words is different in each dataset. We address this problem with novel Prefix Conditioning, a simple yet effective method that helps disentangle dataset biases from visual concepts. This is done by introducing prefix tokens that inform the language encoder of the input data type (e.g., classification vs caption) at training time. Our approach allows the language encoder to learn from both datasets while also tailoring feature extraction to each dataset. Prefix conditioning is generic and can be easily integrated into existing VL pretraining objectives, such as CLIP or UniCL. In experiments, we show that it improves zero-shot image recognition and robustness to image-level distribution shift.
****

Panoptic Lifting for 3D Scene Understanding With Neural Fields

Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Norman Müller, Matthias Nießner, Angela Dai, Peter Kontschieder; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9043-9052

We propose Panoptic Lifting, a novel approach for learning panoptic 3D volumetric representations from images of in-the-wild scenes. Once trained, our model can render color images together with 3D-consistent panoptic segmentation from novel viewpoints. Unlike existing approaches which use 3D input directly or indirectly, our method requires only machine-generated 2D panoptic segmentation masks inferred from a pre-trained network. Our core contribution is a panoptic lifting scheme based on a neural field representation that generates a unified and multi-view consistent, 3D panoptic representation of the scene. To account for inconsistencies of 2D instance identifiers across views, we solve a linear assignment with a cost based on the model's current predictions and the machine-generated segmentation masks, thus enabling us to lift 2D instances to 3D in a consistent way. We further propose and ablate contributions that make our method more robust to noisy, machine-generated labels, including test-time augmentations for confidence estimates, segment consistency loss, bounded segmentation fields, and gradient stopping. Experimental results validate our approach on the challenging Hypersim, Replica, and ScanNet datasets, improving by 8.4, 13.8, and 10.6% in scene-level PQ over state of the art.
****

WeatherStream: Light Transport Automation of Single Image Deweathering

Howard Zhang, Yunhao Ba, Ethan Yang, Varan Mehra, Blake Gella, Akira Suzuki, Arnold Pfahnl, Chethan Chinder Chandrappa, Alex Wong, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13499-13509

Today single image deweathering is arguably more sensitive to the dataset type, rather than the model. We introduce WeatherStream, an automatic pipeline capturing all real-world weather effects (rain, snow, and rain fog degradations), along with their clean image pairs. Previous state-of-the-art methods that have attempted the all-weather removal task train on synthetic pairs, and are thus limited by the Sim2Real domain gap. Recent work has attempted to manually collect time multiplexed pairs, but the use of human labor limits the scale of such a dataset. We introduce a pipeline that uses the power of light-transport physics and a model trained on a small, initial seed dataset to reject approximately 99.6% of unwanted scenes. The pipeline is able to generalize to new scenes and degradations that can, in turn, be used to train existing models just like fully human-labeled data. Training on a dataset collected through this procedure leads to significant improvements on multiple existing weather removal methods on a carefully human-collected test set of real-world weather effects. The dataset and code can be found in the following website: http://visual.ee.ucla.edu/wstream.htm/.

****

Learning To Detect Mirrors From Videos via Dual Correspondences

Jiaying Lin, Xin Tan, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9109-9118

Detecting mirrors from static images has received significant research interest recently. However, detecting mirrors over dynamic scenes is still under-explored due to the lack of a high-quality dataset and an effective method for video mirror detection (VMD). To the best of our knowledge, this is the first work to address the VMD problem from a deep-learning-based perspective. Our observation is that there are often correspondences between the contents inside (reflected) and outside (real) of a mirror, but such correspondences may not always appear in every frame, e.g., due to the change of camera pose. This inspires us to propose a video mirror detection method, named VMD-Net, that can tolerate spatially missing correspondences by considering the mirror correspondences at both the intra-frame level as well as inter-frame level via a dual correspondence module that looks over multiple frames spatially and temporally for correlating correspondences. We further propose a first large-scale dataset for VMD (named VMD-D), which contains 14,987 image frames from 269 videos with corresponding manually annotated masks. Experimental results show that the proposed method outperforms SOTA methods from relevant fields. To enable real-time VMD, our method efficiently utilizes the backbone features by removing the redundant multi-level module design and gets rid of post-processing of the output maps commonly used in existing methods, making it very efficient and practical for real-time video-based applications. Code, dataset, and models are available at https://jiaying.link/cvpr2023-vmd/

****

Single View Scene Scale Estimation Using Scale Field

Byeong-Uk Lee, Jianming Zhang, Yannick Hold-Geoffroy, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21435-21444

In this paper, we propose a single image scale estimation method based on a novel scale field representation. A scale field defines the local pixel-to-metric conversion ratio along the gravity direction on all the ground pixels. This representation resolves the ambiguity in camera parameters, allowing us to use a simple yet effective way to collect scale annotations on arbitrary images from human annotators. By training our model on calibrated panoramic image data and the in-the-wild human annotated data, our single image scene scale estimation network generates robust scale field on a variety of image, which can be utilized in various 3D understanding and scale-aware image editing applications.

****

Learning Semantic-Aware Disentangled Representation for Flexible 3D Human Body E

diting
Xiaokun Sun, Qiao Feng, Xiongzheng Li, Jinsong Zhang, Yu-Kun Lai, Jingyu Yang, Kun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16985-16994
3D human body representation learning has received increasing attention in recent years. However, existing works cannot flexibly, controllably and accurately represent human bodies, limited by coarse semantics and unsatisfactory representation capability, particularly in the absence of supervised data. In this paper, we propose a human body representation with fine-grained semantics and high reconstruction-accuracy in an unsupervised setting. Specifically, we establish a correspondence between latent vectors and geometric measures of body parts by designing a part-aware skeleton-separated decoupling strategy, which facilitates controllable editing of human bodies by modifying the corresponding latent codes. With the help of a bone-guided auto-encoder and an orientation-adaptive weighting strategy, our representation can be trained in an unsupervised manner. With the geometrically meaningful latent space, it can be applied to a wide range of applications, from human body editing to latent code interpolation and shape style transfer. Experimental results on public datasets demonstrate the accurate reconstruction and flexible editing abilities of the proposed method. The code will be available at http://cic.tju.edu.cn/faculty/likun/projects/SemanticHuman.
****
Generating Features With Increased Crop-Related Diversity for Few-Shot Object Detection
Jingyi Xu, Hieu Le, Dimitris Samaras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19713-19722
Two-stage object detectors generate object proposals and classify them to detect objects in images. These proposals often do not perfectly contain the objects but overlap with them in many possible ways, exhibiting great variability in the difficulty levels of the proposals. Training a robust classifier against this crop-related variability requires abundant training data, which is not available in few-shot settings. To mitigate this issue, we propose a novel variational autoencoder (VAE) based data generation model, which is capable of generating data with increased crop-related diversity. The main idea is to transform the latent space such the latent codes with different norms represent different crop-related variations. This allows us to generate features with increased crop-related diversity in difficulty levels by simply varying the latent norm. In particular, each latent code is rescaled such that its norm linearly correlates with the IoU score of the input crop w.r.t. the ground-truth box. Here the IoU score is a proxy that represents the difficulty level of the crop. We train this VAE model on base classes conditioned on the semantic code of each class and then use the trained model to generate features for novel classes. Our experimental results show that our generated features consistently improve state-of-the-art few-shot object detection methods on PASCAL VOC and MS COCO datasets.
****
Towards Scalable Neural Representation for Diverse Videos
Bo He, Xitong Yang, Hanyu Wang, Zuxuan Wu, Hao Chen, Shuaiyi Huang, Yixuan Ren, Ser-Nam Lim, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6132-6142
Implicit neural representations (INR) have gained increasing attention in representing 3D scenes and images, and have been recently applied to encode videos (e.g., NeRV, E-NeRV). While achieving promising results, existing INR-based methods are limited to encoding a handful of short videos (e.g., seven 5-second videos in the UVG dataset) with redundant visual content, leading to a model design that fits individual video frames independently and is not efficiently scalable to a large number of diverse videos. This paper focuses on developing neural representations for a more practical setup -- encoding long and/or a large number of videos with diverse visual content. We first show that instead of dividing videos into small subsets and encoding them with separate models, encoding long and diverse videos jointly with a unified model achieves better compression results. Based on this observation, we propose D-NeRV, a novel neural representation frame

work designed to encode diverse videos by (i) decoupling clip-specific visual content from motion information, (ii) introducing temporal reasoning into the implicit neural network, and (iii) employing the task-oriented flow as intermediate output to reduce spatial redundancies. Our new model largely surpasses NeRV and traditional video compression techniques on UCF101 and UVG datasets on the video compression task. Moreover, when used as an efficient data-loader, D-NeRV achieves 3%-10% higher accuracy than NeRV on action recognition tasks on the UCF101 dataset under the same compression ratios.
****
The Devil Is in the Points: Weakly Semi-Supervised Instance Segmentation via Point-Guided Mask Representation
Beomyoung Kim, Joonhyun Jeong, Dongyoon Han, Sung Ju Hwang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11360-11370
In this paper, we introduce a novel learning scheme named weakly semi-supervised instance segmentation (WSSIS) with point labels for budget-efficient and high-performance instance segmentation. Namely, we consider a dataset setting consisting of a few fully-labeled images and a lot of point-labeled images. Motivated by the main challenge of semi-supervised approaches mainly derives from the trade-off between false-negative and false-positive instance proposals, we propose a method for WSSIS that can effectively leverage the budget-friendly point labels as a powerful weak supervision source to resolve the challenge. Furthermore, to deal with the hard case where the amount of fully-labeled data is extremely limited, we propose a MaskRefineNet that refines noise in rough masks. We conduct extensive experiments on COCO and BDD100K datasets, and the proposed method achieves promising results comparable to those of the fully-supervised model, even with 50% of the fully labeled COCO data (38.8% vs. 39.7%). Moreover, when using as little as 5% of fully labeled COCO data, our method shows significantly superior performance over the state-of-the-art semi-supervised learning method (33.7% vs. 24.9%). The code is available at https://github.com/clovaai/PointWSSIS.
****
Towards Compositional Adversarial Robustness: Generalizing Adversarial Training to Composite Semantic Perturbations
Lei Hsiung, Yun-Yun Tsai, Pin-Yu Chen, Tsung-Yi Ho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24658-24667
Model robustness against adversarial examples of single perturbation type such as the Lp-norm has been widely studied, yet its generalization to more realistic scenarios involving multiple semantic perturbations and their composition remains largely unexplored. In this paper, we first propose a novel method for generating composite adversarial examples. Our method can find the optimal attack composition by utilizing component-wise projected gradient descent and automatic attack-order scheduling. We then propose generalized adversarial training (GAT) to extend model robustness from Lp-ball to composite semantic perturbations, such as the combination of Hue, Saturation, Brightness, Contrast, and Rotation. Results obtained using ImageNet and CIFAR-10 datasets indicate that GAT can be robust not only to all the tested types of a single attack, but also to any combination of such attacks. GAT also outperforms baseline L-infinity-norm bounded adversarial training approaches by a significant margin.
****
Language-Guided Audio-Visual Source Separation via Trimodal Consistency
Reuben Tan, Arijit Ray, Andrea Burns, Bryan A. Plummer, Justin Salamon, Oriol Nieto, Bryan Russell, Kate Saenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10575-10584
We propose a self-supervised approach for learning to perform audio source separation in videos based on natural language queries, using only unlabeled video and audio pairs as training data. A key challenge in this task is learning to associate the linguistic description of a sound-emitting object to its visual features and the corresponding components of the audio waveform, all without access to annotations during training. To overcome this challenge, we adapt off-the-shelf

vision-language foundation models to provide pseudo-target supervision via two novel loss functions and encourage a stronger alignment between the audio, visual and natural language modalities. During inference, our approach can separate sounds given text, video and audio input, or given text and audio input alone. We demonstrate the effectiveness of our self-supervised approach on three audio-visual separation datasets, including MUSIC, SOLOS and AudioSet, where we outperform state-of-the-art strongly supervised approaches despite not using object detectors or text labels during training. Finally, we also include samples of our separated audios in the supplemental for reference.
****
CVT-SLR: Contrastive Visual-Textual Transformation for Sign Language Recognition With Variational Alignment
Jiangbin Zheng, Yile Wang, Cheng Tan, Siyuan Li, Ge Wang, Jun Xia, Yidong Chen, Stan Z. Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23141-23150
Sign language recognition (SLR) is a weakly supervised task that annotates sign videos as textual glosses. Recent studies show that insufficient training caused by the lack of large-scale available sign datasets becomes the main bottleneck for SLR. Most SLR works thereby adopt pretrained visual modules and develop two mainstream solutions. The multi-stream architectures extend multi-cue visual features, yielding the current SOTA performances but requiring complex designs and might introduce potential noise. Alternatively, the advanced single-cue SLR frameworks using explicit cross-modal alignment between visual and textual modalities are simple and effective, potentially competitive with the multi-cue framework. In this work, we propose a novel contrastive visual-textual transformation for SLR, CVT-SLR, to fully explore the pretrained knowledge of both the visual and language modalities. Based on the single-cue cross-modal alignment framework, we propose a variational autoencoder (VAE) for pretrained contextual knowledge while introducing the complete pretrained language module. The VAE implicitly aligns visual and textual modalities while benefiting from pretrained contextual knowledge as the traditional contextual module. Meanwhile, a contrastive cross-modal alignment algorithm is designed to explicitly enhance the consistency constraints. Extensive experiments on public datasets (PHOENIX-2014 and PHOENIX-2014T) demonstrate that our proposed CVT-SLR consistently outperforms existing single-cue methods and even outperforms SOTA multi-cue methods.
****
DynaMask: Dynamic Mask Selection for Instance Segmentation
Ruihuang Li, Chenhang He, Shuai Li, Yabin Zhang, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11279-11288
The representative instance segmentation methods mostly segment different object instances with a mask of the fixed resolution, e.g., 28x 28 grid. However, a low-resolution mask loses rich details, while a high-resolution mask incurs quadratic computation overhead. It is a challenging task to predict the optimal binary mask for each instance. In this paper, we propose to dynamically select suitable masks for different object proposals. First, a dual-level Feature Pyramid Network (FPN) with adaptive feature aggregation is developed to gradually increase the mask grid resolution, ensuring high-quality segmentation of objects. Specifically, an efficient region-level top-down path (r-FPN) is introduced to incorporate complementary contextual and detailed information from different stages of image-level FPN (i-FPN). Then, to alleviate the increase of computation and memory costs caused by using large masks, we develop a Mask Switch Module (MSM) with negligible computational cost to select the most suitable mask resolution for each instance, achieving high efficiency while maintaining high segmentation accuracy. Without bells and whistles, the proposed method, namely DynaMask, brings consistent and noticeable performance improvements over other state-of-the-arts at a moderate computation overhead. The source code: https://github.com/lslrh/DynaMask.
****
Paint by Example: Exemplar-Based Image Editing With Diffusion Models

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Ch
en, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 18381-18391
Language-guided image editing has achieved great success recently. In this paper
, we investigate exemplar-guided image editing for more precise control. We achi
eve this goal by leveraging self-supervised training to disentangle and re-organ
ize the source image and the exemplar. However, the naive approach will cause ob
vious fusing artifacts. We carefully analyze it and propose an information bottl
eneck and strong augmentations to avoid the trivial solution of directly copying
 and pasting the exemplar image. Meanwhile, to ensure the controllability of the
 editing process, we design an arbitrary shape mask for the exemplar image and l
everage the classifier-free guidance to increase the similarity to the exemplar
image. The whole framework involves a single forward of the diffusion model with
out any iterative optimization. We demonstrate that our method achieves an impre
ssive performance and enables controllable editing on in-the-wild images with hi
gh fidelity.
****
Ego-Body Pose Estimation via Ego-Head Pose Estimation
Jiaman Li, Karen Liu, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2023, pp. 17142-17151
Estimating 3D human motion from an egocentric video sequence plays a critical ro
le in human behavior understanding and has various applications in VR/AR. Howeve
r, naively learning a mapping between egocentric videos and human motions is cha
llenging, because the user's body is often unobserved by the front-facing camera
 placed on the head of the user. In addition, collecting large-scale, high-quali
ty datasets with paired egocentric videos and 3D human motions requires accurate
 motion capture devices, which often limit the variety of scenes in the videos t
o lab-like environments. To eliminate the need for paired egocentric video and h
uman motions, we propose a new method, Ego-Body Pose Estimation via Ego-Head Pos
e Estimation (EgoEgo), which decomposes the problem into two stages, connected b
y the head motion as an intermediate representation. EgoEgo first integrates SLA
M and a learning approach to estimate accurate head motion. Subsequently, levera
ging the estimated head pose as input, EgoEgo utilizes conditional diffusion to
generate multiple plausible full-body motions. This disentanglement of head and
body pose eliminates the need for training datasets with paired egocentric video
s and 3D human motion, enabling us to leverage large-scale egocentric video data
sets and motion capture datasets separately. Moreover, for systematic benchmarki
ng, we develop a synthetic dataset, AMASS-Replica-Ego-Syn (ARES), with paired eg
ocentric videos and human motion. On both ARES and real data, our EgoEgo model p
erforms significantly better than the current state-of-the-art methods.
****
SAP-DETR: Bridging the Gap Between Salient Points and Queries-Based Transformer
Detector for Fast Model Convergency
Yang Liu, Yao Zhang, Yixin Wang, Yang Zhang, Jiang Tian, Zhongchao Shi, Jianping
 Fan, Zhiqiang He; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2023, pp. 15539-15547
Recently, the dominant DETR-based approaches apply central-concept spatial prior
 to accelerating Transformer detector convergency. These methods gradually refin
e the reference points to the center of target objects and imbue object queries
with the updated central reference information for spatially conditional attenti
on. However, centralizing reference points may severely deteriorate queries' sal
iency and confuse detectors due to the indiscriminative spatial prior. To bridge
 the gap between the reference points of salient queries and Transformer detecto
rs, we propose SAlient Point-based DETR (SAP-DETR) by treating object detection
as a transformation from salient points to instance objects. In SAP-DETR, we exp
licitly initialize a query-specific reference point for each object query, gradu
ally aggregate them into an instance object, and then predict the distance from
each side of the bounding box to these points. By rapidly attending to query-spe
cific reference region and other conditional extreme regions from the image feat
ures, SAP-DETR can effectively bridge the gap between the salient point and the

query-based Transformer detector with a significant convergency speed. Our exten sive experiments have demonstrated that SAP-DETR achieves 1.4 times convergency speed with competitive performance. Under the standard training scheme, SAP-DETR stably promotes the SOTA approaches by 1.0 AP. Based on ResNet-DC-101, SAP-DET R achieves 46.9 AP. The code will be released at https://github.com/liuyang-ict/ SAP-DETR.
****
GD-MAE: Generative Decoder for MAE Pre-Training on LiDAR Point Clouds
Honghui Yang, Tong He, Jiaheng Liu, Hua Chen, Boxi Wu, Binbin Lin, Xiaofei He, W anli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2023, pp. 9403-9414
Despite the tremendous progress of Masked Autoencoders (MAE) in developing visio n tasks such as image and video, exploring MAE in large-scale 3D point clouds re mains challenging due to the inherent irregularity. In contrast to previous 3D M AE frameworks, which either design a complex decoder to infer masked information from maintained regions or adopt sophisticated masking strategies, we instead p ropose a much simpler paradigm. The core idea is to apply a Generative Decoder f or MAE (GD-MAE) to automatically merges the surrounding context to restore the m asked geometric knowledge in a hierarchical fusion manner. In doing so, our appr oach is free from introducing the heuristic design of decoders and enjoys the fl exibility of exploring various masking strategies. The corresponding part costs less than 12% latency compared with conventional methods, while achieving better performance. We demonstrate the efficacy of the proposed method on several larg e-scale benchmarks: Waymo, KITTI, and ONCE. Consistent improvement on downstream detection tasks illustrates strong robustness and generalization capability. No t only our method reveals state-of-the-art results, but remarkably, we achieve c omparable accuracy even with 20% of the labeled data on the Waymo dataset. Code will be released.
****
Towards Robust Tampered Text Detection in Document Image: New Dataset and New So lution
Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, Lia nwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5937-5946
Recently, tampered text detection in document image has attracted increasingly a ttention due to its essential role on information security. However, detecting v isually consistent tampered text in photographed document images is still a main challenge. In this paper, we propose a novel framework to capture more fine-gra ined clues in complex scenarios for tampered text detection, termed as Document Tampering Detector (DTD), which consists of a Frequency Perception Head (FPH) to compensate the deficiencies caused by the inconspicuous visual features, and a Multi-view Iterative Decoder (MID) for fully utilizing the information of featur es in different scales. In addition, we design a new training paradigm, termed a s Curriculum Learning for Tampering Detection (CLTD), which can address the conf usion during the training procedure and thus to improve the robustness for image compression and the ability to generalize. To further facilitate the tampered t ext detection in document images, we construct a large-scale document image data set, termed as DocTamper, which contains 170,000 document images of various type s. Experiments demonstrate that our proposed DTD outperforms previous state-of-t he-art by 9.2%, 26.3% and 12.3% in terms of F-measure on the DocTamper testing s et, and the cross-domain testing sets of DocTamper-FCD and DocTamper-SCD, respec tively. Codes and dataset will be available at https://github.com/qcf-568/DocTam per.
****
Learning Rotation-Equivariant Features for Visual Correspondence
Jongmin Lee, Byungjin Kim, Seungwook Kim, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21887-2 1897
Extracting discriminative local features that are invariant to imaging variation s is an integral part of establishing correspondences between images. In this wo

rk, we introduce a self-supervised learning framework to extract discriminative rotation-invariant descriptors using group-equivariant CNNs. Thanks to employing group-equivariant CNNs, our method effectively learns to obtain rotation-equivariant features and their orientations explicitly, without having to perform sophisticated data augmentations. The resultant features and their orientations are further processed by group aligning, a novel invariant mapping technique that shifts the group-equivariant features by their orientations along the group dimension. Our group aligning technique achieves rotation-invariance without any collapse of the group dimension and thus eschews loss of discriminability. The proposed method is trained end-to-end in a self-supervised manner, where we use an orientation alignment loss for the orientation estimation and a contrastive descriptor loss for robust local descriptors to geometric/photometric variations. Our method demonstrates state-of-the-art matching accuracy among existing rotation-invariant descriptors under varying rotation and also shows competitive results when transferred to the task of keypoint matching and camera pose estimation.
****
DexArt: Benchmarking Generalizable Dexterous Manipulation With Articulated Objects
Chen Bao, Helin Xu, Yuzhe Qin, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21190-21200
To enable general-purpose robots, we will require the robot to operate daily articulated objects as humans do. Current robot manipulation has heavily relied on using a parallel gripper, which restricts the robot to a limited set of objects. On the other hand, operating with a multi-finger robot hand will allow better approximation to human behavior and enable the robot to operate on diverse articulated objects. To this end, we propose a new benchmark called DexArt, which involves Dexterous manipulation with Articulated objects in a physical simulator. In our benchmark, we define multiple complex manipulation tasks, and the robot hand will need to manipulate diverse articulated objects within each task. Our main focus is to evaluate the generalizability of the learned policy on unseen articulated objects. This is very challenging given the high degrees of freedom of both hands and objects. We use Reinforcement Learning with 3D representation learning to achieve generalization. Through extensive studies, we provide new insights into how 3D representation learning affects decision making in RL with 3D point cloud inputs. More details can be found at https://www.chenbao.tech/dexart/.
****
DeSTSeg: Segmentation Guided Denoising Student-Teacher for Anomaly Detection
Xuan Zhang, Shiyu Li, Xi Li, Ping Huang, Jiulong Shan, Ting Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3914-3923
Visual anomaly detection, an important problem in computer vision, is usually formulated as a one-class classification and segmentation task. The student-teacher (S-T) framework has proved to be effective in solving this challenge. However, previous works based on S-T only empirically applied constraints on normal data and fused multi-level information. In this study, we propose an improved model called DeSTSeg, which integrates a pre-trained teacher network, a denoising student encoder-decoder, and a segmentation network into one framework. First, to strengthen the constraints on anomalous data, we introduce a denoising procedure that allows the student network to learn more robust representations. From synthetically corrupted normal images, we train the student network to match the teacher network feature of the same images without corruption. Second, to fuse the multi-level S-T features adaptively, we train a segmentation network with rich supervision from synthetic anomaly masks, achieving a substantial performance improvement. Experiments on the industrial inspection benchmark dataset demonstrate that our method achieves state-of-the-art performance, 98.6% on image-level AUC, 75.8% on pixel-level average precision, and 76.4% on instance-level average precision.
****
Neural Rate Estimator and Unsupervised Learning for Efficient Distributed Image Analytics in Split-DNN Models

Nilesh Ahuja, Parual Datta, Bhavya Kanzariya, V. Srinivasa Somayazulu, Omesh Tickoo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2022-2030
Thanks to advances in computer vision and AI, there has been a large growth in the demand for cloud-based visual analytics in which images captured by a low-powered edge device are transmitted to the cloud for analytics. Use of conventional codecs (JPEG, MPEG, HEVC, etc.) for compressing such data introduces artifacts that can seriously degrade the performance of the downstream analytic tasks. Split-DNN computing has emerged as a paradigm to address such usages, in which a DNN is partitioned into a client-side portion and a server side portion. Low-complexity neural networks called 'bottleneck units' are introduced at the split point to transform the intermediate layer features into a lower-dimensional representation better suited for compression and transmission. Optimizing the pipeline for both compression and task-performance requires high-quality estimates of the information-theoretic rate of the intermediate features. Most works on compression for image analytics use heuristic approaches to estimate the rate, leading to suboptimal performance. We propose a high-quality 'neural rate-estimator' to address this gap. We interpret the lower-dimensional bottleneck output as a latent representation of the intermediate feature and cast the rate-distortion optimization problem as one of training an equivalent variational auto-encoder with an appropriate loss function. We show that this leads to improved rate-distortion outcomes. We further show that replacing supervised loss terms (such as cross-entropy loss) by distillation-based losses in a teacher-student framework allows for unsupervised training of bottleneck units without the need for explicit training labels. This makes our method very attractive for real world deployments where access to labeled training data is difficult or expensive. We demonstrate that our method outperforms several state-of-the-art methods by obtaining improved task accuracy at lower bitrates on image classification and semantic segmentation tasks.
****
Object Pop-Up: Can We Infer 3D Objects and Their Poses From Human Interactions Alone?
Ilya A. Petrov, Riccardo Marin, Julian Chibane, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4726-4736
The intimate entanglement between objects affordances and human poses is of large interest, among others, for behavioural sciences, cognitive psychology, and Computer Vision communities. In recent years, the latter has developed several object-centric approaches: starting from items, learning pipelines synthesizing human poses and dynamics in a realistic way, satisfying both geometrical and functional expectations. However, the inverse perspective is significantly less explored: Can we infer 3D objects and their poses from human interactions alone? Our investigation follows this direction, showing that a generic 3D human point cloud is enough to pop up an unobserved object, even when the user is just imitating a functionality (e.g., looking through a binocular) without involving a tangible counterpart. We validate our method qualitatively and quantitatively, with synthetic data and sequences acquired for the task, showing applicability for XR/VR.
****
VoP: Text-Video Co-Operative Prompt Tuning for Cross-Modal Retrieval
Siteng Huang, Biao Gong, Yulin Pan, Jianwen Jiang, Yiliang Lv, Yuyuan Li, Donglin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6565-6574
Many recent studies leverage the pre-trained CLIP for text-video cross-modal retrieval by tuning the backbone with additional heavy modules, which not only brings huge computational burdens with much more parameters, but also leads to the knowledge forgetting from upstream models. In this work, we propose the VoP: Text-Video Co-operative Prompt Tuning for efficient tuning on the text-video retrieval task. The proposed VoP is an end-to-end framework with both video & text prompts introducing, which can be regarded as a powerful baseline with only 0.1% trainable parameters. Further, based on the spatio-temporal characteristics of vide

os, we develop three novel video prompt mechanisms to improve the performance with different scales of trainable parameters. The basic idea of the VoP enhancement is to model the frame position, frame context, and layer function with specific trainable prompts, respectively. Extensive experiments show that compared to full fine-tuning, the enhanced VoP achieves a 1.4% average R@1 gain across five text-video retrieval benchmarks with 6x less parameter overhead. The code will be available at https://github.com/bighuang624/VoP.
****
Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR
Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6873-6883
This paper advances the fine-grained sketch-based image retrieval (FG-SBIR) literature by putting forward a strong baseline that overshoots prior state-of-the art by 11%. This is not via complicated design though, but by addressing two critical issues facing the community (i) the gold standard triplet loss does not enforce holistic latent space geometry, and (ii) there are never enough sketches to train a high accuracy model. For the former, we propose a simple modification to the standard triplet loss, that explicitly enforces separation amongst photos/sketch instances. For the latter, we put forward a novel knowledge distillation module can leverage photo data for model training. Both modules are then plugged into a novel plug-n-playable training paradigm that allows for more stable training. More specifically, for (i) we employ an intra-modal triplet loss amongst sketches to bring sketches of the same instance closer from others, and one more amongst photos to push away different photo instances while bringing closer a structurally augmented version of the same photo (offering a gain of 4-6%). To tackle (ii), we first pre-train a teacher on the large set of unlabelled photos over the aforementioned intra-modal photo triplet loss. Then we distill the contextual similarity present amongst the instances in the teacher's embedding space to that in the student's embedding space, by matching the distribution over inter-feature distances of respective samples in both embedding spaces (delivering a further gain of 4-5%). Apart from outperforming prior arts significantly, our model also yields satisfactory results on generalising to new classes. Project page: https://aneeshan95.github.io/Sketch_PVT/
****
You Do Not Need Additional Priors or Regularizers in Retinex-Based Low-Light Image Enhancement
Huiyuan Fu, Wenkai Zheng, Xiangyu Meng, Xin Wang, Chuanming Wang, Huadong Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18125-18134
Images captured in low-light conditions often suffer from significant quality degradation. Recent works have built a large variety of deep Retinex-based networks to enhance low-light images. The Retinex-based methods require decomposing the image into reflectance and illumination components, which is a highly ill-posed problem and there is no available ground truth. Previous works addressed this problem by imposing some additional priors or regularizers. However, finding an effective prior or regularizer that can be applied in various scenes is challenging, and the performance of the model suffers from too many additional constraints. We propose a contrastive learning method and a self-knowledge distillation method that allow training our Retinex-based model for Retinex decomposition without elaborate hand-crafted regularization functions. Rather than estimating reflectance and illuminance images and representing the final images as their element-wise products as in previous works, our regularizer-free Retinex decomposition and synthesis network (RFR) extracts reflectance and illuminance features and synthesizes them end-to-end. In addition, we propose a loss function for contrastive learning and a progressive learning strategy for self-knowledge distillation. Extensive experimental results demonstrate that our proposed methods can achieve superior performance compared with state-of-the-art approaches.
****
PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification

Meike Nauta, Jörg Schlötterer, Maurice van Keulen, Christin Seifert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2744-2753
Interpretable methods based on prototypical patches recognize various components in an image in order to explain their reasoning to humans. However, existing prototype-based methods can learn prototypes that are not in line with human visual perception, i.e., the same prototype can refer to different concepts in the real world, making interpretation not intuitive. Driven by the principle of explainability-by-design, we introduce PIP-Net (Patch-based Intuitive Prototypes Network): an interpretable image classification model that learns prototypical parts in a self-supervised fashion which correlate better with human vision. PIP-Net can be interpreted as a sparse scoring sheet where the presence of a prototypical part in an image adds evidence for a class. The model can also abstain from a decision for out-of-distribution data by saying "I haven't seen this before". We only use image-level labels and do not rely on any part annotations. PIP-Net is globally interpretable since the set of learned prototypes shows the entire reasoning of the model. A smaller local explanation locates the relevant prototypes in one image. We show that our prototypes correlate with ground-truth object parts, indicating that PIP-Net closes the "semantic gap" between latent space and pixel space. Hence, our PIP-Net with interpretable prototypes enables users to interpret the decision making process in an intuitive, faithful and semantically meaningful way. Code is available at https://github.com/M-Nauta/PIPNet.
****
SCADE: NeRFs from Space Carving With Ambiguity-Aware Depth Estimates
Mikaela Angelina Uy, Ricardo Martin-Brualla, Leonidas Guibas, Ke Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16518-16527
Neural radiance fields (NeRFs) have enabled high fidelity 3D reconstruction from multiple 2D input views. However, a well-known drawback of NeRFs is the less-than-ideal performance under a small number of views, due to insufficient constraints enforced by volumetric rendering. To address this issue, we introduce SCADE, a novel technique that improves NeRF reconstruction quality on sparse, unconstrained input views for in-the-wild indoor scenes. To constrain NeRF reconstruction, we leverage geometric priors in the form of per-view depth estimates produced with state-of-the-art monocular depth estimation models, which can generalize across scenes. A key challenge is that monocular depth estimation is an ill-posed problem, with inherent ambiguities. To handle this issue, we propose a new method that learns to predict, for each view, a continuous, multimodal distribution of depth estimates using conditional Implicit Maximum Likelihood Estimation (cIMLE). In order to disambiguate exploiting multiple views, we introduce an original space carving loss that guides the NeRF representation to fuse multiple hypothesized depth maps from each view and distill from them a common geometry that is consistent with all views. Experiments show that our approach enables higher fidelity novel view synthesis from sparse views. Our project page can be found at https://scade-spacecarving-nerfs.github.io.
****
Re-Thinking Model Inversion Attacks Against Deep Neural Networks
Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16384-16393
Model inversion (MI) attacks aim to infer and reconstruct private training data by abusing access to a model. MI attacks have raised concerns about the leaking of sensitive information (e.g. private face images used in training a face recognition system). Recently, several algorithms for MI have been proposed to improve the attack performance. In this work, we revisit MI, study two fundamental issues pertaining to all state-of-the-art (SOTA) MI algorithms, and propose solutions to these issues which lead to a significant boost in attack performance for all SOTA MI. In particular, our contributions are two-fold: 1) We analyze the optimization objective of SOTA MI algorithms, argue that the objective is sub-optimal for achieving MI, and propose an improved optimization objective that boosts

attack performance significantly. 2) We analyze "MI overfitting", show that it would prevent reconstructed images from learning semantics of training data, and propose a novel "model augmentation" idea to overcome this issue. Our proposed solutions are simple and improve all SOTA MI attack accuracy significantly. E.g., in the standard CelebA benchmark, our solutions improve accuracy by 11.8% and achieve for the first time over 90% attack accuracy. Our findings demonstrate that there is a clear risk of leaking sensitive information from deep learning models. We urge serious consideration to be given to the privacy implications. Our code, demo, and models are available at https://ngoc-nguyen-0.github.io/re-thinking_model_inversion_attacks/
****
1% VS 100%: Parameter-Efficient Low Rank Adapter for Dense Predictions
Dongshuo Yin, Yiran Yang, Zhechao Wang, Hongfeng Yu, Kaiwen Wei, Xian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20116-20126
Fine-tuning large-scale pre-trained vision models to downstream tasks is a standard technique for achieving state-of-the-art performance on computer vision benchmarks. However, fine-tuning the whole model with millions of parameters is inefficient as it requires storing a same-sized new model copy for each task. In this work, we propose LoRand, a method for fine-tuning large-scale vision models with a better trade-off between task performance and the number of trainable parameters. LoRand generates tiny adapter structures with low-rank synthesis while keeping the original backbone parameters fixed, resulting in high parameter sharing. To demonstrate LoRand's effectiveness, we implement extensive experiments on object detection, semantic segmentation, and instance segmentation tasks. By only training a small percentage (1% to 3%) of the pre-trained backbone parameters, LoRand achieves comparable performance to standard fine-tuning on COCO and ADE20K and outperforms fine-tuning in low-resource PASCAL VOC dataset.
****
ResFormer: Scaling ViTs With Multi-Resolution Training
Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22721-22731
Vision Transformers (ViTs) have achieved overwhelming success, yet they suffer from vulnerable resolution scalability, i.e., the performance drops drastically when presented with input resolutions that are unseen during training. We introduce, ResFormer, a framework that is built upon the seminal idea of multi-resolution training for improved performance on a wide spectrum of, mostly unseen, testing resolutions. In particular, ResFormer operates on replicated images of different resolutions and enforces a scale consistency loss to engage interactive information across different scales. More importantly, to alternate among varying resolutions effectively, especially novel ones in testing, we propose a global-local positional embedding strategy that changes smoothly conditioned on input sizes. We conduct extensive experiments for image classification on ImageNet. The results provide strong quantitative evidence that ResFormer has promising scaling abilities towards a wide range of resolutions. For instance, ResFormer- B-MR achieves a Top-1 accuracy of 75.86% and 81.72% when evaluated on relatively low and high resolutions respectively (i.e., 96 and 640), which are 48% and 7.49% better than DeiT-B. We also demonstrate, moreover, ResFormer is flexible and can be easily extended to semantic segmentation, object detection and video action recognition.
****
You Need Multiple Exiting: Dynamic Early Exiting for Accelerating Unified Vision Language Model
Shengkun Tang, Yaqing Wang, Zhenglun Kong, Tianchi Zhang, Yao Li, Caiwen Ding, Yanzhi Wang, Yi Liang, Dongkuan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10781-10791
Large-scale transformer models bring significant improvements for various downstream vision language tasks with a unified architecture. The performance improvements come with increasing model size, resulting in slow inference speed and incr

eased cost for severing. While some certain predictions benefit from the full co
mplexity of the large-scale model, not all of input need the same amount of comp
utation to conduct, potentially leading to computation resource waste. To handle
 this challenge, early exiting is proposed to adaptively allocate computational
power in term of input complexity to improve inference efficiency. The existing
early exiting strategies usually adopt output confidence based on intermediate l
ayers as a proxy of input complexity to incur the decision of skipping following
 layers. However, such strategies cannot apply to encoder in the widely-used uni
fied architecture with both encoder and decoder due to difficulty of output conf
idence estimation in the encoder. It is suboptimal in term of saving computation
 power to ignore the early exiting in encoder component. To handle this challeng
e, we propose a novel early exiting strategy for unified visual language models,
 which allows dynamically skip the layers in encoder and decoder simultaneously
in term of input layer-wise similarities with multiple times of early exiting, n
amely MuE. By decomposing the image and text modalities in the encoder, MuE is f
lexible and can skip different layers in term of modalities, advancing the infer
ence efficiency while minimizing performance drop. Experiments on the SNLI-VE an
d MS COCO datasets show that the proposed approach MuE can reduce inference time
 by up to 50% and 40% while maintaining 99% and 96% performance respectively.
****
CloSET: Modeling Clothed Humans on Continuous Surface With Explicit Template Dec
omposition
Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Y
andong Guo, Yebin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision
 and Pattern Recognition (CVPR), 2023, pp. 501-511
Creating animatable avatars from static scans requires the modeling of clothing
deformations in different poses. Existing learning-based methods typically add p
ose-dependent deformations upon a minimally-clothed mesh template or a learned i
mplicit template, which have limitations in capturing details or hinder end-to-e
nd learning. In this paper, we revisit point-based solutions and propose to deco
mpose explicit garment-related templates and then add pose-dependent wrinkles to
 them. In this way, the clothing deformations are disentangled such that the pos
e-dependent wrinkles can be better learned and applied to unseen poses. Addition
ally, to tackle the seam artifact issues in recent state-of-the-art point-based
methods, we propose to learn point features on a body surface, which establishes
 a continuous and compact feature space to capture the fine-grained and pose-dep
endent clothing geometry. To facilitate the research in this field, we also intr
oduce a high-quality scan dataset of humans in real-world clothing. Our approach
 is validated on two existing datasets and our newly introduced dataset, showing
 better clothing deformation results in unseen poses. The project page with code
 and dataset can be found at https://www.liuyebin.com/closet.
****
BUOL: A Bottom-Up Framework With Occupancy-Aware Lifting for Panoptic 3D Scene R
econstruction From a Single Image
Tao Chu, Pan Zhang, Qiong Liu, Jiaqi Wang; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4937-4946
Understanding and modeling the 3D scene from a single image is a practical probl
em. A recent advance proposes a panoptic 3D scene reconstruction task that perfo
rms both 3D reconstruction and 3D panoptic segmentation from a single image. Alt
hough having made substantial progress, recent works only focus on top-down appr
oaches that fill 2D instances into 3D voxels according to estimated depth, which
 hinders their performance by two ambiguities. (1) instance-channel ambiguity: T
he variable ids of instances in each scene lead to ambiguity during filling voxe
l channels with 2D information, confusing the following 3D refinement. (2) voxel
-reconstruction ambiguity: 2D-to-3D lifting with estimated single view depth onl
y propagates 2D information onto the surface of 3D regions, leading to ambiguity
 during the reconstruction of regions behind the frontal view surface. In this p
aper, we propose BUOL, a Bottom-Up framework with Occupancy-aware Lifting to add
ress the two issues for panoptic 3D scene reconstruction from a single image. Fo
r instance-channel ambiguity, a bottom-up framework lifts 2D information to 3D v

oxels based on deterministic semantic assignments rather than arbitrary instance id assignments. The 3D voxels are then refined and grouped into 3D instances according to the predicted 2D instance centers. For voxel-reconstruction ambiguity, the estimated multi-plane occupancy is leveraged together with depth to fill the whole regions of things and stuff. Our method shows a tremendous performance advantage over state-of-the-art methods on synthetic dataset 3D-Front and real-world dataset Matterport3D, respectively. Code and models will be released.
****

Hierarchical Video-Moment Retrieval and Step-Captioning
Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, Mohit Bansal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23056-23065
There is growing interest in searching for information from large video corpora. Prior works have studied relevant tasks, such as text-based video retrieval, moment retrieval, video summarization, and video captioning in isolation, without an end-to-end setup that can jointly search from video corpora and generate summaries. Such an end-to-end setup would allow for many interesting applications, e.g., a text-based search that finds a relevant video from a video corpus, extracts the most relevant moment from that video, and segments the moment into important steps with captions. To address this, we present the HiREST (HIerarchical REtrieval and STep-captioning) dataset and propose a new benchmark that covers hierarchical information retrieval and visual/textual stepwise summarization from an instructional video corpus. HiREST consists of 3.4K text-video pairs from an instructional video dataset, where 1.1K videos have annotations of moment spans relevant to text query and breakdown of each moment into key instruction steps with caption and timestamps (totaling 8.6K step captions). Our hierarchical benchmark consists of video retrieval, moment retrieval, and two novel moment segmentation and step captioning tasks. In moment segmentation, models break down a video moment into instruction steps and identify start-end boundaries. In step captioning, models generate a textual summary for each step. We also present starting point task-specific and end-to-end joint baseline models for our new benchmark. While the baseline models show some promising results, there still exists large room for future improvement by the community.
****

PROB: Probabilistic Objectness for Open World Object Detection
Orr Zohar, Kuan-Chieh Wang, Serena Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11444-11453
Open World Object Detection (OWOD) is a new and challenging computer vision task that bridges the gap between classic object detection (OD) benchmarks and object detection in the real world. In addition to detecting and classifying seen/labeled objects, OWOD algorithms are expected to detect novel/unknown objects - which can be classified and incrementally learned. In standard OD, object proposals not overlapping with a labeled object are automatically classified as background. Therefore, simply applying OD methods to OWOD fails as unknown objects would be predicted as background. The challenge of detecting unknown objects stems from the lack of supervision in distinguishing unknown objects and background object proposals. Previous OWOD methods have attempted to overcome this issue by generating supervision using pseudo-labeling - however, unknown object detection has remained low. Probabilistic/generative models may provide a solution for this challenge. Herein, we introduce a novel probabilistic framework for objectness estimation, where we alternate between probability distribution estimation and objectness likelihood maximization of known objects in the embedded feature space - ultimately allowing us to estimate the objectness probability of different proposals. The resulting Probabilistic Objectness transformer-based open-world detector, PROB, integrates our framework into traditional object detection models, adapting them for the open-world setting. Comprehensive experiments on OWOD benchmarks show that PROB outperforms all existing OWOD methods in both unknown object detection (2x unknown recall) and known object detection (mAP). Our code is available at https://github.com/orrzohar/PROB.
****

PD-Quant: Post-Training Quantization Based on Prediction Difference Metric
Jiawei Liu, Lin Niu, Zhihang Yuan, Dawei Yang, Xinggang Wang, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24427-24437
Post-training quantization (PTQ) is a neural network compression technique that converts a full-precision model into a quantized model using lower-precision data types. Although it can help reduce the size and computational cost of deep neural networks, it can also introduce quantization noise and reduce prediction accuracy, especially in extremely low-bit settings. How to determine the appropriate quantization parameters (e.g., scaling factors and rounding of weights) is the main problem facing now. Existing methods attempt to determine these parameters by minimize the distance between features before and after quantization, but such an approach only considers local information and may not result in the most optimal quantization parameters. We analyze this issue and propose PD-Quant, a method that addresses this limitation by considering global information. It determines the quantization parameters by using the information of differences between network prediction before and after quantization. In addition, PD-Quant can alleviate the overfitting problem in PTQ caused by the small number of calibration sets by adjusting the distribution of activations. Experiments show that PD-Quant leads to better quantization parameters and improves the prediction accuracy of quantized models, especially in low-bit settings. For example, PD-Quant pushes the accuracy of ResNet-18 up to 53.14% and RegNetX-600MF up to 40.67% in weight 2-bit activation 2-bit. The code is released at https://github.com/hustvl/PD-Quant.
****
AUNet: Learning Relations Between Action Units for Face Forgery Detection
Weiming Bai, Yufan Liu, Zhipeng Zhang, Bing Li, Weiming Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24709-24719
Face forgery detection becomes increasingly crucial due to the serious security issues caused by face manipulation techniques. Recent studies in deepfake detection have yielded promising results when the training and testing face forgeries are from the same domain. However, the problem remains challenging when one tries to generalize the detector to forgeries created by unseen methods during training. Observing that face manipulation may alter the relation between different facial action units (AU), we propose the Action Units Relation Learning framework to improve the generality of forgery detection. In specific, it consists of the Action Units Relation Transformer (ART) and the Tampered AU Prediction (TAP). The ART constructs the relation between different AUs with AU-agnostic Branch and AU-specific Branch, which complement each other and work together to exploit forgery clues. In the Tampered AU Prediction, we tamper AU-related regions at the image level and develop challenging pseudo samples at the feature level. The model is then trained to predict the tampered AU regions with the generated location-specific supervision. Experimental results demonstrate that our method can achieve state-of-the-art performance in both the in-dataset and cross-dataset evaluations.
****
SparseFusion: Distilling View-Conditioned Diffusion for 3D Reconstruction
Zhizhuo Zhou, Shubham Tulsiani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12588-12597
We propose SparseFusion, a sparse view 3D reconstruction approach that unifies recent advances in neural rendering and probabilistic image generation. Existing approaches typically build on neural rendering with re-projected features but fail to generate unseen regions or handle uncertainty under large viewpoint changes. Alternate methods treat this as a (probabilistic) 2D synthesis task, and while they can generate plausible 2D images, they do not infer a consistent underlying 3D. However, we find that this trade-off between 3D consistency and probabilistic image generation does not need to exist. In fact, we show that geometric consistency and generative inference can be complementary in a mode seeking behavior. By distilling a 3D consistent scene representation from a view-conditioned l

atent diffusion model, we are able to recover a plausible 3D representation whose renderings are both accurate and realistic. We evaluate our approach across 51 categories in the CO3D dataset and show that it outperforms existing methods, in both distortion and perception metrics, for sparse view novel view synthesis.
****
PolyFormer: Referring Image Segmentation As Sequential Polygon Generation
Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, R. Manmatha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18653-18663
In this work, instead of directly predicting the pixel-level segmentation masks, the problem of referring image segmentation is formulated as sequential polygon generation, and the predicted polygons can be later converted into segmentation masks. This is enabled by a new sequence-to-sequence framework, Polygon Transformer (PolyFormer), which takes a sequence of image patches and text query tokens as input, and outputs a sequence of polygon vertices autoregressively. For more accurate geometric localization, we propose a regression-based decoder, which predicts the precise floating-point coordinates directly, without any coordinate quantization error. In the experiments, PolyFormer outperforms the prior art by a clear margin, e.g., 5.40% and 4.52% absolute improvements on the challenging RefCOCO+ and RefCOCOg datasets. It also shows strong generalization ability when evaluated on the referring video segmentation task without fine-tuning, e.g., achieving competitive 61.5% J&F on the Ref-DAVIS17 dataset.
****
Seeing What You Miss: Vision-Language Pre-Training With Semantic Completion Learning
Yatai Ji, Rongcheng Tu, Jie Jiang, Weijie Kong, Chengfei Cai, Wenzhe Zhao, Hongfa Wang, Yujiu Yang, Wei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6789-6798
Cross-modal alignment is essential for vision-language pre-training (VLP) models to learn the correct corresponding information across different modalities. For this purpose, inspired by the success of masked language modeling (MLM) tasks in the NLP pre-training area, numerous masked modeling tasks have been proposed for VLP to further promote cross-modal interactions. The core idea of previous masked modeling tasks is to focus on reconstructing the masked tokens based on visible context for learning local-to-local alignment. However, most of them pay little attention to the global semantic features generated for the masked data, resulting in a limited cross-modal alignment ability of global representations. Therefore, in this paper, we propose a novel Semantic Completion Learning (SCL) task, complementary to existing masked modeling tasks, to facilitate global-to-local alignment. Specifically, the SCL task complements the missing semantics of masked data by capturing the corresponding information from the other modality, promoting learning more representative global features which have a great impact on the performance of downstream tasks. Moreover, we present a flexible vision encoder, which enables our model to perform image-text and video-text multimodal tasks simultaneously. Experimental results show that our proposed method obtains state-of-the-art performance on various vision-language benchmarks, such as visual question answering, image-text retrieval, and video-text retrieval.
****
Interactive Segmentation As Gaussion Process Classification
Minghao Zhou, Hong Wang, Qian Zhao, Yuexiang Li, Yawen Huang, Deyu Meng, Yefeng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19488-19497
Click-based interactive segmentation (IS) aims to extract the target objects under user interaction. For this task, most of the current deep learning (DL)-based methods mainly follow the general pipelines of semantic segmentation. Albeit achieving promising performance, they do not fully and explicitly utilize and propagate the click information, inevitably leading to unsatisfactory segmentation results, even at clicked points. Against this issue, in this paper, we propose to formulate the IS task as a Gaussian process (GP)-based pixel-wise binary classification model on each image. To solve this model, we utilize amortized variatio

nal inference to approximate the intractable GP posterior in a data-driven manne
r and then decouple the approximated GP posterior into double space forms for ef
ficient sampling with linear complexity. Then, we correspondingly construct a GP
 classification framework, named GPCIS, which is integrated with the deep kernel
 learning mechanism for more flexibility. The main specificities of the proposed
 GPCIS lie in: 1) Under the explicit guidance of the derived GP posterior, the i
nformation contained in clicks can be finely propagated to the entire image and
then boost the segmentation; 2) The accuracy of predictions at clicks has good t
heoretical support. These merits of GPCIS as well as its good generality and hig
h efficiency are substantiated by comprehensive experiments on several benchmark
s, as compared with representative methods both quantitatively and qualitatively
. Codes will be released at https://github.com/zmhhmz/GPCIS_CVPR2023.
****
Differentiable Shadow Mapping for Efficient Inverse Graphics
Markus Worchel, Marc Alexa; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2023, pp. 142-153
We show how shadows can be efficiently generated in differentiable rendering of
triangle meshes. Our central observation is that pre-filtered shadow mapping, a
technique for approximating shadows based on rendering from the perspective of a
 light, can be combined with existing differentiable rasterizers to yield differ
entiable visibility information. We demonstrate at several inverse graphics prob
lems that differentiable shadow maps are orders of magnitude faster than differe
ntiable light transport simulation with similar accuracy -- while differentiable
 rasterization without shadows often fails to converge.
****
Dynamic Focus-Aware Positional Queries for Semantic Segmentation
Haoyu He, Jianfei Cai, Zizheng Pan, Jing Liu, Jing Zhang, Dacheng Tao, Bohan Zhu
ang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2023, pp. 11299-11308
The DETR-like segmentors have underpinned the most recent breakthroughs in seman
tic segmentation, which end-to-end train a set of queries representing the class
 prototypes or target segments. Recently, masked attention is proposed to restri
ct each query to only attend to the foreground regions predicted by the precedin
g decoder block for easier optimization. Although promising, it relies on the le
arnable parameterized positional queries which tend to encode the dataset statis
tics, leading to inaccurate localization for distinct individual queries. In thi
s paper, we propose a simple yet effective query design for semantic segmentatio
n termed Dynamic Focus-aware Positional Queries (DFPQ), which dynamically genera
tes positional queries conditioned on the cross-attention scores from the preced
ing decoder block and the positional encodings for the corresponding image featu
res, simultaneously. Therefore, our DFPQ preserves rich localization information
 for the target segments and provides accurate and fine-grained positional prior
s. In addition, we propose to efficiently deal with high-resolution cross-attent
ion by only aggregating the contextual tokens based on the low-resolution cross-
attention scores to perform local relation aggregation. Extensive experiments on
 ADE20K and Cityscapes show that with the two modifications on Mask2former, our
framework achieves SOTA performance and outperforms Mask2former by clear margins
 of 1.1%, 1.9%, and 1.1% single-scale mIoU with ResNet-50, Swin-T, and Swin-B ba
ckbones on the ADE20K validation set, respectively. Source code is available at
https://github.com/ziplab/FASeg.
****
A Practical Stereo Depth System for Smart Glasses
Jialiang Wang, Daniel Scharstein, Akash Bapat, Kevin Blackburn-Matzen, Matthew Y
u, Jonathan Lehman, Suhib Alsisan, Yanghan Wang, Sam Tsai, Jan-Michael Frahm, Zi
jian He, Peter Vajda, Michael F. Cohen, Matt Uyttendaele; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21
498-21507
We present the design of a productionized end-to-end stereo depth sensing system
 that does pre-processing, online stereo rectification, and stereo depth estimat
ion with a fallback to monocular depth estimation when rectification is unreliab

le. The output of our depth sensing system is then used in a novel view generation pipeline to create 3D computational photography effects using point-of-view images captured by smart glasses. All these steps are executed on-device on the stringent compute budget of a mobile phone, and because we expect the users can use a wide range of smartphones, our design needs to be general and cannot be dependent on a particular hardware or ML accelerator such as a smartphone GPU. Although each of these steps is well studied, a description of a practical system is still lacking. For such a system, all these steps need to work in tandem with one another and fallback gracefully on failures within the system or less than ideal input data. We show how we handle unforeseen changes to calibration, e.g., due to heat, robustly support depth estimation in the wild, and still abide by the memory and latency constraints required for a smooth user experience. We show that our trained models are fast, and run in less than 1s on a six-year-old Samsung Galaxy S8 phone's CPU. Our models generalize well to unseen data and achieve good results on Middlebury and in-the-wild images captured from the smart glasses.
****
Understanding and Constructing Latent Modality Structures in Multi-Modal Representation Learning
Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, Trishul Chilimbi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7661-7671
Contrastive loss has been increasingly used in learning representations from multiple modalities. In the limit, the nature of the contrastive loss encourages modalities to exactly match each other in the latent space. Yet it remains an open question how the modality alignment affects the downstream task performance. In this paper, based on an information-theoretic argument, we first prove that exact modality alignment is sub-optimal in general for downstream prediction tasks. Hence we advocate that the key of better performance lies in meaningful latent modality structures instead of perfect modality alignment. To this end, we propose three general approaches to construct latent modality structures. Specifically, we design 1) a deep feature separation loss for intra-modality regularization; 2) a Brownian-bridge loss for inter-modality regularization; and 3) a geometric consistency loss for both intra- and inter-modality regularization. Extensive experiments are conducted on two popular multi-modal representation learning frameworks: the CLIP-based two-tower model and the ALBEF-based fusion model. We test our model on a variety of tasks including zero/few-shot image classification, image-text retrieval, visual question answering, visual reasoning, and visual entailment. Our method achieves consistent improvements over existing methods, demonstrating the effectiveness and generalizability of our proposed approach on latent modality structure regularization.
****
PointConvFormer: Revenge of the Point-Based Convolution
Wenxuan Wu, Li Fuxin, Qi Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21802-21813
We introduce PointConvFormer, a novel building block for point cloud based deep network architectures. Inspired by generalization theory, PointConvFormer combines ideas from point convolution, where filter weights are only based on relative position, and Transformers which utilize feature-based attention. In PointConvFormer, attention computed from feature difference between points in the neighborhood is used to modify the convolutional weights at each point. Hence, we preserved the invariances from point convolution, whereas attention helps to select relevant points in the neighborhood for convolution. We experiment on both semantic segmentation and scene flow estimation tasks on point clouds with multiple datasets including ScanNet, SemanticKitti, FlyingThings3D and KITTI. Our results show that PointConvFormer substantially outperforms classic convolutions, regular transformers, and voxelized sparse convolution approaches with much smaller and faster networks. Visualizations show that PointConvFormer performs similarly to convolution on flat areas, whereas the neighborhood selection effect is stronger on object boundaries, showing that it has got the best of both worlds. The code

will be available with the final version.
****
Instant Volumetric Head Avatars
Wojciech Zielonka, Timo Bolkart, Justus Thies; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4574-4584
We present Instant Volumetric Head Avatars (INSTA), a novel approach for reconstructing photo-realistic digital avatars instantaneously. INSTA models a dynamic neural radiance field based on neural graphics primitives embedded around a parametric face model. Our pipeline is trained on a single monocular RGB portrait video that observes the subject under different expressions and views. While state-of-the-art methods take up to several days to train an avatar, our method can reconstruct a digital avatar in less than 10 minutes on modern GPU hardware, which is orders of magnitude faster than previous solutions. In addition, it allows for the interactive rendering of novel poses and expressions. By leveraging the geometry prior of the underlying parametric face model, we demonstrate that INSTA extrapolates to unseen poses. In quantitative and qualitative studies on various subjects, INSTA outperforms state-of-the-art methods regarding rendering quality and training time. Project website: https://zielon.github.io/insta/
****
HARP: Personalized Hand Reconstruction From a Monocular RGB Video
Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, Siyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12802-12813
We present HARP (HAnd Reconstruction and Personalization), a personalized hand avatar creation approach that takes a short monocular RGB video of a human hand as input and reconstructs a faithful hand avatar exhibiting a high-fidelity appearance and geometry. In contrast to the major trend of neural implicit representations, HARP models a hand with a mesh-based parametric hand model, a vertex displacement map, a normal map, and an albedo without any neural components. The explicit nature of our representation enables a truly scalable, robust, and efficient approach to hand avatar creation as validated by our experiments. HARP is optimized via gradient descent from a short sequence captured by a hand-held mobile phone and can be directly used in AR/VR applications with real-time rendering capability. To enable this, we carefully design and implement a shadow-aware differentiable rendering scheme that is robust to high degree articulations and self-shadowing regularly present in hand motions, as well as challenging lighting conditions. It also generalizes to unseen poses and novel viewpoints, producing photo-realistic renderings of hand animations. Furthermore, the learned HARP representation can be used for improving 3D hand pose estimation quality in challenging viewpoints. The key advantages of HARP are validated by the in-depth analyses on appearance reconstruction, novel view and novel pose synthesis, and 3D hand pose refinement. It is an AR/VR-ready personalized hand representation that shows superior fidelity and scalability.
****
Variational Distribution Learning for Unsupervised Text-to-Image Generation
Minsoo Kang, Doyup Lee, Jiseob Kim, Saehoon Kim, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23380-23389
We propose a text-to-image generation algorithm based on deep neural networks when text captions for images are unavailable during training. In this work, instead of simply generating pseudo-ground-truth sentences of training images using existing image captioning methods, we employ a pretrained CLIP model, which is capable of properly aligning embeddings of images and corresponding texts in a joint space and, consequently, works well on zero-shot recognition tasks. We optimize a text-to-image generation model by maximizing the data log-likelihood conditioned on pairs of image-text CLIP embeddings. To better align data in the two domains, we employ a principled way based on a variational inference, which efficiently estimates an approximate posterior of the hidden text embedding given an image and its CLIP feature. Experimental results validate that the proposed framework outperforms existing approaches by large margins under unsupervised and sem

i-supervised text-to-image generation settings.
****
MetaMix: Towards Corruption-Robust Continual Learning With Temporally Self-Adaptive Data Transformation
Zhenyi Wang, Li Shen, Donglin Zhan, Qiuling Suo, Yanjun Zhu, Tiehang Duan, Mingchen Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24521-24531
Continual Learning (CL) has achieved rapid progress in recent years. However, it is still largely unknown how to determine whether a CL model is trustworthy and how to foster its trustworthiness. This work focuses on evaluating and improving the robustness to corruptions of existing CL models. Our empirical evaluation results show that existing state-of-the-art (SOTA) CL models are particularly vulnerable to various data corruptions during testing. To make them trustworthy and robust to corruptions deployed in safety-critical scenarios, we propose a meta-learning framework of self-adaptive data augmentation to tackle the corruption robustness in CL. The proposed framework, MetaMix, learns to augment and mix data, automatically transforming the new task data or memory data. It directly optimizes the generalization performance against data corruptions during training. To evaluate the corruption robustness of our proposed approach, we construct several CL corruption datasets with different levels of severity. We perform comprehensive experiments on both task- and class-continual learning. Extensive experiments demonstrate the effectiveness of our proposed method compared to SOTA baselines.
****
Ultra-High Resolution Segmentation With Ultra-Rich Context: A Novel Benchmark
Deyi Ji, Feng Zhao, Hongtao Lu, Mingyuan Tao, Jieping Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23621-23630
With the increasing interest and rapid development of methods for Ultra-High Resolution (UHR) segmentation, a large-scale benchmark covering a wide range of scenes with full fine-grained dense annotations is urgently needed to facilitate the field. To this end, the URUR dataset is introduced, in the meaning of Ultra-High Resolution dataset with Ultra-Rich Context. As the name suggests, URUR contains amounts of images with high enough resolution (3,008 images of size 5,120x5,120), a wide range of complex scenes (from 63 cities), rich-enough context (1 million instances with 8 categories) and fine-grained annotations (about 80 billion manually annotated pixels), which is far superior to all the existing UHR datasets including DeepGlobe, Inria Aerial, UDD, etc.. Moreover, we also propose WSDNet, a more efficient and effective framework for UHR segmentation especially with ultra-rich context. Specifically, multi-level Discrete Wavelet Transform (DWT) is naturally integrated to release computation burden while preserve more spatial details, along with a Wavelet Smooth Loss (WSL) to reconstruct original structured context and texture with a smooth constrain. Experiments on several UHR datasets demonstrate its state-of-the-art performance. The dataset is available at https://github.com/jankyee/URUR.
****
DART: Diversify-Aggregate-Repeat Training Improves Generalization of Neural Networks
Samyak Jain, Sravanti Addepalli, Pawan Kumar Sahu, Priyam Dey, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16048-16059
Generalization of Neural Networks is crucial for deploying them safely in the real world. Common training strategies to improve generalization involve the use of data augmentations, ensembling and model averaging. In this work, we first establish a surprisingly simple but strong benchmark for generalization which utilizes diverse augmentations within a training minibatch, and show that this can learn a more balanced distribution of features. Further, we propose Diversify-Aggregate-Repeat Training (DART) strategy that first trains diverse models using different augmentations (or domains) to explore the loss basin, and further Aggregates their weights to combine their expertise and obtain improved generalization.

We find that Repeating the step of Aggregation throughout training improves the
overall optimization trajectory and also ensures that the individual models hav
e sufficiently low loss barrier to obtain improved generalization on combining t
hem. We theoretically justify the proposed approach and show that it indeed gene
ralizes better. In addition to improvements in In-Domain generalization, we demo
nstrate SOTA performance on the Domain Generalization benchmarks in the popular
DomainBed framework as well. Our method is generic and can easily be integrated
with several base training algorithms to achieve performance gains. Our code is
available here: https://github.com/val-iisc/DART.
****
Cross-Domain Image Captioning With Discriminative Finetuning
Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonir
ina, Francesca Franzon, Marco Baroni; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6935-6944
Neural captioners are typically trained to mimic human-generated references with
out optimizing for any specific communication goal, leading to problems such as
the generation of vague captions. In this paper, we show that fine-tuning an out
-of-the-box neural captioner with a self-supervised discriminative communication
objective helps to recover a plain, visually descriptive language that is more
informative about image contents. Given a target image, the system must learn to
produce a description that enables an out-of-the-box text-conditioned image ret
riever to identify such image among a set of candidates. We experiment with the
popular ClipCap captioner, also replicating the main results with BLIP. In terms
of similarity to ground-truth human descriptions, the captions emerging from di
scriminative finetuning lag slightly behind those generated by the non-finetuned
model, when the latter is trained and tested on the same caption dataset. Howev
er, when the model is used without further tuning to generate captions for out-o
f-domain datasets, our discriminatively-finetuned captioner generates descriptio
ns that resemble human references more than those produced by the same captioner
without finetuning. We further show that, on the Conceptual Captions dataset, d
iscriminatively finetuned captions are more helpful than either vanilla ClipCap
captions or ground-truth captions for human annotators tasked with an image disc
rimination task.
****
Accelerating Vision-Language Pretraining With Free Language Modeling
Teng Wang, Yixiao Ge, Feng Zheng, Ran Cheng, Ying Shan, Xiaohu Qie, Ping Luo; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2023, pp. 23161-23170
The state of the arts in vision-language pretraining (VLP) achieves exemplary pe
rformance but suffers from high training costs resulting from slow convergence a
nd long training time, especially on large-scale web datasets. An essential obst
acle to training efficiency lies in the entangled prediction rate (percentage of
tokens for reconstruction) and corruption rate (percentage of corrupted tokens)
in masked language modeling (MLM), that is, a proper corruption rate is achieve
d at the cost of a large portion of output tokens being excluded from prediction
loss. To accelerate the convergence of VLP, we propose a new pretraining task,
namely, free language modeling (FLM), that enables a 100% prediction rate with a
rbitrary corruption rates. FLM successfully frees the prediction rate from the t
ie-up with the corruption rate while allowing the corruption spans to be customi
zed for each token to be predicted. FLM-trained models are encouraged to learn b
etter and faster given the same GPU time by exploiting bidirectional contexts mo
re flexibly. Extensive experiments show FLM could achieve an impressive 2.5x pre
training time reduction in comparison to the MLM-based methods, while keeping co
mpetitive performance on both vision-language understanding and generation tasks
.
****
Efficient Mask Correction for Click-Based Interactive Image Segmentation
Fei Du, Jianlong Yuan, Zhibin Wang, Fan Wang; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22773-22782
The goal of click-based interactive image segmentation is to extract target mask

s with the input of positive/negative clicks. Every time a new click is placed, existing methods run the whole segmentation network to obtain a corrected mask, which is inefficient since several clicks may be needed to reach satisfactory accuracy. To this end, we propose an efficient method to correct the mask with a lightweight mask correction network. The whole network remains a low computational cost from the second click, even if we have a large backbone. However, a simple correction network with limited capacity is not likely to achieve comparable performance with a classic segmentation network. Thus, we propose a click-guided self-attention module and a click-guided correlation module to effectively exploits the click information to boost performance. First, several templates are selected based on the semantic similarity with click features. Then the self-attention module propagates the template information to other pixels, while the correlation module directly uses the templates to obtain target outlines. With the efficient architecture and two click-guided modules, our method shows preferable performance and efficiency compared to existing methods. The code will be released at https://github.com/feiaxyt/EMC-Click.
****
DBARF: Deep Bundle-Adjusting Generalizable Neural Radiance Fields
Yu Chen, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24-34
Recent works such as BARF and GARF can bundle adjust camera poses with neural radiance fields (NeRF) which is based on coordinate-MLPs. Despite the impressive results, these methods cannot be applied to Generalizable NeRFs (GeNeRFs) which require image feature extractions that are often based on more complicated 3D CNN or transformer architectures. In this work, we first analyze the difficulties of jointly optimizing camera poses with GeNeRFs, and then further propose our DBARF to tackle these issues. Our DBARF which bundle adjusts camera poses by taking a cost feature map as an implicit cost function can be jointly trained with GeNeRFs in a self-supervised manner. Unlike BARF and its follow-up works, which can only be applied to per-scene optimized NeRFs and need accurate initial camera poses with the exception of forward-facing scenes, our method can generalize across scenes and does not require any good initialization. Experiments show the effectiveness and generalization ability of our DBARF when evaluated on real-world datasets. Our code is available at https://aibluefisher.github.io/dbarf.
****
EvShutter: Transforming Events for Unconstrained Rolling Shutter Correction
Julius Erbach, Stepan Tulyakov, Patricia Vitoria, Alfredo Bochicchio, Yuanyou Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13904-13913
Widely used Rolling Shutter (RS) CMOS sensors capture high resolution images at the expense of introducing distortions and artifacts in the presence of motion. In such situations, RS distortion correction algorithms are critical. Recent methods rely on a constant velocity assumption and require multiple frames to predict the dense displacement field. In this work, we introduce a new method, called Eventful Shutter (EvShutter), that corrects RS artifacts using a single RGB image and event information with high temporal resolution. The method firstly removes blur using a novel flow-based deblurring module and then compensates RS using a double encoder hourglass network. In contrast to previous methods, it does not rely on a constant velocity assumption and uses a simple architecture thanks to an event transformation dedicated to RS, called Filter and Flip (FnF), that transforms input events to encode only the changes between GS and RS images. To evaluate the proposed method and facilitate future research, we collect the first dataset with real events and high-quality RS images with optional blur, called RS-ERGB. We generate the RS images from GS images using a newly proposed simulator based on adaptive interpolation. The simulator permits the use of inexpensive cameras with long exposure to capture high-quality GS images. We show that on this realistic dataset the proposed method outperforms the state-of-the-art image- and event-based methods by 9.16 dB and 0.75 dB respectively in terms of PSNR and an improvement of 23% and 21% in LPIPS.
****

Graphics Capsule: Learning Hierarchical 3D Face Representations From 2D Images
Chang Yu, Xiangyu Zhu, Xiaomei Zhang, Zhaoxiang Zhang, Zhen Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20981-20990
The function of constructing the hierarchy of objects is important to the visual process of the human brain. Previous studies have successfully adopted capsule networks to decompose the digits and faces into parts in an unsupervised manner to investigate the similar perception mechanism of neural networks. However, their descriptions are restricted to the 2D space, limiting their capacities to imitate the intrinsic 3D perception ability of humans. In this paper, we propose an Inverse Graphics Capsule Network (IGC-Net) to learn the hierarchical 3D face representations from large-scale unlabeled images. The core of IGC-Net is a new type of capsule, named graphics capsule, which represents 3D primitives with interpretable parameters in computer graphics (CG), including depth, albedo, and 3D pose. Specifically, IGC-Net first decomposes the objects into a set of semantic-consistent part-level descriptions and then assembles them into object-level descriptions to build the hierarchy. The learned graphics capsules reveal how the neural networks, oriented at visual perception, understand faces as a hierarchy of 3D models. Besides, the discovered parts can be deployed to the unsupervised face segmentation task to evaluate the semantic consistency of our method. Moreover, the part-level descriptions with explicit physical meanings provide insight into the face analysis that originally runs in a black box, such as the importance of shape and texture for face recognition. Experiments on CelebA, BP4D, and Multi-PIE demonstrate the characteristics of our IGC-Net.
****
Connecting the Dots: Floorplan Reconstruction Using Two-Level Queries
Yuanwen Yue, Theodora Kontogianni, Konrad Schindler, Francis Engelmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 845-854
We address 2D floorplan reconstruction from 3D scans. Existing approaches typically employ heuristically designed multi-stage pipelines. Instead, we formulate floorplan reconstruction as a single-stage structured prediction task: find a variable-size set of polygons, which in turn are variable-length sequences of ordered vertices. To solve it we develop a novel Transformer architecture that generates polygons of multiple rooms in parallel, in a holistic manner without hand-crafted intermediate stages. The model features two-level queries for polygons and corners, and includes polygon matching to make the network end-to-end trainable. Our method achieves a new state-of-the-art for two challenging datasets, Structured3D and SceneCAD, along with significantly faster inference than previous methods. Moreover, it can readily be extended to predict additional information, i.e., semantic room types and architectural elements like doors and windows. Our code and models are available at: https://github.com/ywyue/RoomFormer.
****
Analyzing and Diagnosing Pose Estimation With Attributions
Qiyuan He, Linlin Yang, Kerui Gu, Qiuxia Lin, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4821-4830
We present Pose Integrated Gradient (PoseIG), the first interpretability technique designed for pose estimation. We extend the concept of integrated gradients for pose estimation to generate pixel-level attribution maps. To enable comparison across different pose frameworks, we unify different pose outputs into a common output space, along with a likelihood approximation function for gradient back-propagation. To complement the qualitative insight from the attribution maps, we propose three indices for quantitative analysis. With these tools, we systematically compare different pose estimation frameworks to understand the impacts of network design, backbone and auxiliary tasks. Our analysis reveals an interesting shortcut of the knuckles (MCP joints) for hand pose estimation and an under-explored inversion error for keypoints in body pose estimation. Project page: https://qy-h00.github.io/poseig/.
****

Ambiguity-Resistant Semi-Supervised Learning for Dense Object Detection
Chang Liu, Weiming Zhang, Xiangru Lin, Wei Zhang, Xiao Tan, Junyu Han, Xiaomao Li, Errui Ding, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15579-15588
With basic Semi-Supervised Object Detection (SSOD) techniques, one-stage detectors generally obtain limited promotions compared with two-stage clusters. We experimentally find that the root lies in two kinds of ambiguities: (1) Selection ambiguity that selected pseudo labels are less accurate, since classification scores cannot properly represent the localization quality. (2) Assignment ambiguity that samples are matched with improper labels in pseudo-label assignment, as the strategy is misguided by missed objects and inaccurate pseudo boxes. To tackle these problems, we propose a Ambiguity-Resistant Semi-supervised Learning (ARSL) for one-stage detectors. Specifically, to alleviate the selection ambiguity, Joint-Confidence Estimation (JCE) is proposed to jointly quantifies the classification and localization quality of pseudo labels. As for the assignment ambiguity, Task-Separation Assignment (TSA) is introduced to assign labels based on pixel-level predictions rather than unreliable pseudo boxes. It employs a 'divide-and-conquer' strategy and separately exploits positives for the classification and localization task, which is more robust to the assignment ambiguity. Comprehensive experiments demonstrate that ARSL effectively mitigates the ambiguities and achieves state-of-the-art SSOD performance on MS COCO and PASCAL VOC. Codes can be found at https://github.com/PaddlePaddle/PaddleDetection.
****
Scalable, Detailed and Mask-Free Universal Photometric Stereo
Satoshi Ikehata; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13198-13207
In this paper, we introduce SDM-UniPS, a groundbreaking Scalable, Detailed, Mask-free, and Universal Photometric Stereo network. Our approach can recover astonishingly intricate surface normal maps, rivaling the quality of 3D scanners, even when images are captured under unknown, spatially-varying lighting conditions in uncontrolled environments. We have extended previous universal photometric stereo networks to extract spatial-light features, utilizing all available information in high-resolution input images and accounting for non-local interactions among surface points. Moreover, we present a new synthetic training dataset that encompasses a diverse range of shapes, materials, and illumination scenarios found in real-world scenes. Through extensive evaluation, we demonstrate that our method not only surpasses calibrated, lighting-specific techniques on public benchmarks, but also excels with a significantly smaller number of input images even without object masks.
****
Towards High-Quality and Efficient Video Super-Resolution via Spatial-Temporal Data Overfitting
Gen Li, Jie Ji, Minghai Qin, Wei Niu, Bin Ren, Fatemeh Afghah, Linke Guo, Xiaolong Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10259-10269
As deep convolutional neural networks (DNNs) are widely used in various fields of computer vision, leveraging the overfitting ability of the DNN to achieve video resolution upscaling has become a new trend in the modern video delivery system. By dividing videos into chunks and overfitting each chunk with a super-resolution model, the server encodes videos before transmitting them to the clients, thus achieving better video quality and transmission efficiency. However, a large number of chunks are expected to ensure good overfitting quality, which substantially increases the storage and consumes more bandwidth resources for data transmission. On the other hand, decreasing the number of chunks through training optimization techniques usually requires high model capacity, which significantly slows down execution speed. To reconcile such, we propose a novel method for high-quality and efficient video resolution upscaling tasks, which leverages the spatial-temporal information to accurately divide video into chunks, thus keeping the number of chunks as well as the model size to a minimum. Additionally, we advance our method into a single overfitting model by a data-aware joint training

technique, which further reduces the storage requirement with negligible quality drop. We deploy our proposed overfitting models on an off-the-shelf mobile phone, and experimental results show that our method achieves real-time video super-resolution with high video quality. Compared with the state-of-the-art, our method achieves 28 fps streaming speed with 41.60 PSNR, which is 14 times faster and 2.29 dB better in the live video resolution upscaling tasks.
****
Make-a-Story: Visual Memory Conditioned Consistent Story Generation
Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, Leonid Sigal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2493-2502
There has been a recent explosion of impressive generative models that can produce high quality images (or videos) conditioned on text descriptions. However, all such approaches rely on conditional sentences that contain unambiguous descriptions of scenes and main actors in them. Therefore employing such models for more complex task of story visualization, where naturally references and co-references exist, and one requires to reason about when to maintain consistency of actors and backgrounds across frames/scenes, and when not to, based on story progression, remains a challenge. In this work, we address the aforementioned challenges and propose a novel autoregressive diffusion-based framework with a visual memory module that implicitly captures the actor and background context across the generated frames. Sentence-conditioned soft attention over the memories enables effective reference resolution and learns to maintain scene and actor consistency when needed. To validate the effectiveness of our approach, we extend the MUGEN dataset and introduce additional characters, backgrounds and referencing in multi-sentence storylines. Our experiments for story generation on the MUGEN, the PororoSV and the FlintstonesSV dataset show that our method not only outperforms prior state-of-the-art in generating frames with high visual quality, which are consistent with the story, but also models appropriate correspondences between the characters and the background.
****
BiFormer: Vision Transformer With Bi-Level Routing Attention
Lei Zhu, Xinjiang Wang, Zhanghan Ke, Wayne Zhang, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10323-10333
As the core building block of vision transformers, attention is a powerful tool to capture long-range dependency. However, such power comes at a cost: it incurs a huge computation burden and heavy memory footprint as pairwise token interaction across all spatial locations is computed. A series of works attempt to alleviate this problem by introducing handcrafted and content-agnostic sparsity into attention, such as restricting the attention operation to be inside local windows, axial stripes, or dilated windows. In contrast to these approaches, we propose a novel dynamic sparse attention via bi-level routing to enable a more flexible allocation of computations with content awareness. Specifically, for a query, irrelevant key-value pairs are first filtered out at a coarse region level, and then fine-grained token-to-token attention is applied in the union of remaining candidate regions (i.e., routed regions). We provide a simple yet effective implementation of the proposed bi-level routing attention, which utilizes the sparsity to save both computation and memory while involving only GPU-friendly dense matrix multiplications. Built with the proposed bi-level routing attention, a new general vision transformer, named BiFormer, is then presented. As BiFormer attends to a small subset of relevant tokens in a query-adaptive manner without distraction from other irrelevant ones, it enjoys both good performance and high computational efficiency, especially in dense prediction tasks. Empirical results across several computer vision tasks such as image classification, object detection, and semantic segmentation verify the effectiveness of our design. Code is available at https://github.com/rayleizhu/BiFormer.
****
Masked Autoencoders Enable Efficient Knowledge Distillers
Yutong Bai, Zeyu Wang, Junfei Xiao, Chen Wei, Huiyu Wang, Alan L. Yuille, Yuyin

Zhou, Cihang Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24256-24265
This paper studies the potential of distilling knowledge from pre-trained models, especially Masked Autoencoders. Our approach is simple: in addition to optimizing the pixel reconstruction loss on masked inputs, we minimize the distance between the intermediate feature map of the teacher model and that of the student model. This design leads to a computationally efficient knowledge distillation framework, given 1) only a small visible subset of patches is used, and 2) the (cumbersome) teacher model only needs to be partially executed, i.e., forward propagate inputs through the first few layers, for obtaining intermediate feature maps. Compared to directly distilling fine-tuned models, distilling pre-trained models substantially improves downstream performance. For example, by distilling the knowledge from an MAE pre-trained ViT-L into a ViT-B, our method achieves 84.0% ImageNet top-1 accuracy, outperforming the baseline of directly distilling a fine-tuned ViT-L by 1.2%. More intriguingly, our method can robustly distill knowledge from teacher models even with extremely high masking ratios: e.g., with 95% masking ratio where merely TEN patches are visible during distillation, our ViT-B competitively attains a top-1 ImageNet accuracy of 83.6%; surprisingly, it can still secure 82.4% top-1 ImageNet accuracy by aggressively training with just FOUR visible patches (98% masking ratio). The code will be made publicly available.
****
TinyMIM: An Empirical Study of Distilling MIM Pre-Trained Models
Sucheng Ren, Fangyun Wei, Zheng Zhang, Han Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3687-3697
Masked image modeling (MIM) performs strongly in pre-training large vision Transformers (ViTs). However, small models that are critical for real-world applications cannot or only marginally benefit from this pre-training approach. In this paper, we explore distillation techniques to transfer the success of large MIM-based pre-trained models to smaller ones. We systematically study different options in the distillation framework, including distilling targets, losses, input, network regularization, sequential distillation, etc, revealing that: 1) Distilling token relations is more effective than CLS token- and feature-based distillation; 2) An intermediate layer of the teacher network as target perform better than that using the last layer when the depth of the student mismatches that of the teacher; 3) Weak regularization is preferred; etc. With these findings, we achieve significant fine-tuning accuracy improvements over the scratch MIM pre-training on ImageNet-1K classification, using all the ViT-Tiny, ViT-Small, and ViT-base models, with +4.2%/+2.4%/+1.4% gains, respectively. Our TinyMIM model of base size achieves 52.2 mIoU in AE20K semantic segmentation, which is +4.1 higher than the MAE baseline. Our TinyMIM model of tiny size achieves 79.6% top-1 accuracy on ImageNet-1K image classification, which sets a new record for small vision models of the same size and computation budget. This strong performance suggests an alternative way for developing small vision Transformer models, that is, by exploring better training methods rather than introducing inductive biases into architectures as in most previous works. Code is available at https://github.com/OliverRensu/TinyMIM.
****
Persistent Nature: A Generative Model of Unbounded 3D Worlds
Lucy Chai, Richard Tucker, Zhengqi Li, Phillip Isola, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20863-20874
Despite increasingly realistic image quality, recent 3D image generative models often operate on 3D volumes of fixed extent with limited camera motions. We investigate the task of unconditionally synthesizing unbounded nature scenes, enabling arbitrarily large camera motion while maintaining a persistent 3D world model. Our scene representation consists of an extendable, planar scene layout grid, which can be rendered from arbitrary camera poses via a 3D decoder and volume rendering, and a panoramic skydome. Based on this representation, we learn a generative world model solely from single-view internet photos. Our method enables si

mulating long flights through 3D landscapes, while maintaining global scene cons
istency---for instance, returning to the starting point yields the same view of
the scene. Our approach enables scene extrapolation beyond the fixed bounds of c
urrent 3D generative models, while also supporting a persistent, camera-independ
ent world representation that stands in contrast to auto-regressive 3D predictio
n models. Our project page: https://chail.github.io/persistent-nature/.
****
OneFormer: One Transformer To Rule Universal Image Segmentation
Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, Humphrey Shi;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2023, pp. 2989-2998
Universal Image Segmentation is not a new concept.Past attempts to unify image s
egmentation include scene parsing, panoptic segmentation, and, more recently, ne
w panoptic architectures. However, such panoptic architectures do not truly unif
y image segmentation because they need to be trained individually on the semanti
c, instance, or panoptic segmentation to achieve the best performance. Ideally,
a truly universal framework should be trained only once and achieve SOTA perform
ance across all three image segmentation tasks. To that end, we propose OneForme
r, a universal image segmentation framework that unifies segmentation with a mul
ti-task train-once design. We first propose a task-conditioned joint training st
rategy that enables training on ground truths of each domain (semantic, instance
, and panoptic segmentation) within a single multi-task training process. Second
ly, we introduce a task token to condition our model on the task at hand, making
 our model task-dynamic to support multi-task training and inference. Thirdly, w
e propose using a query-text contrastive loss during training to establish bette
r inter-task and inter-class distinctions. Notably, our single OneFormer model o
utperforms specialized Mask2Former models across all three segmentation tasks on
 ADE20k, Cityscapes, and COCO, despite the latter being trained on each task ind
ividually. We believe OneFormer is a significant step towards making image segme
ntation more universal and accessible.
****
Hierarchical Neural Memory Network for Low Latency Event Processing
Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, Ken Sakurada; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 22867-22876
This paper proposes a low latency neural network architecture for event-based de
nse prediction tasks. Conventional architectures encode entire scene contents at
 a fixed rate regardless of their temporal characteristics. Instead, the propose
d network encodes contents at a proper temporal scale depending on its movement
speed. We achieve this by constructing temporal hierarchy using stacked latent m
emories that operate at different rates. Given low latency event steams, the mul
ti-level memories gradually extract dynamic to static scene contents by propagat
ing information from the fast to the slow memory modules. The architecture not o
nly reduces the redundancy of conventional architectures but also exploits long-
term dependencies. Furthermore, an attention-based event representation efficien
tly encodes sparse event streams into the memory cells. We conduct extensive eva
luations on three event-based dense prediction tasks, where the proposed approac
h outperforms the existing methods on accuracy and latency, while demonstrating
effective event and image fusion capabilities. The code is available at https://
hamarh.github.io/hmnet/
****
Finding Geometric Models by Clustering in the Consensus Space
Daniel Barath, Denys Rozumnyi, Ivan Eichhardt, Levente Hajder, Jiri Matas; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2023, pp. 5414-5424
We propose a new algorithm for finding an unknown number of geometric models, e.
g., homographies. The problem is formalized as finding dominant model instances
progressively without forming crisp point-to-model assignments. Dominant instanc
es are found via a RANSAC-like sampling and a consolidation process driven by a
model quality function considering previously proposed instances. New ones are f

ound by clustering in the consensus space. This new formulation leads to a simpl e iterative algorithm with state-of-the-art accuracy while running in real-time on a number of vision problems -- at least two orders of magnitude faster than t he competitors on two-view motion estimation. Also, we propose a deterministic s ampler reflecting the fact that real-world data tend to form spatially coherent structures. The sampler returns connected components in a progressively densifie d neighborhood-graph. We present a number of applications where the use of multi ple geometric models improves accuracy. These include pose estimation from multi ple generalized homographies; trajectory estimation of fast-moving objects; and we also propose a way of using multiple homographies in global SfM algorithms. S ource code: https://github.com/danini/clustering-in-consensus-space.
****
Leapfrog Diffusion Model for Stochastic Trajectory Prediction
Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, Yanfeng Wang; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 55 17-5526
To model the indeterminacy of human behaviors, stochastic trajectory prediction requires a sophisticated multi-modal distribution of future trajectories. Emergi ng diffusion models have revealed their tremendous representation capacities in numerous generation tasks, showing potential for stochastic trajectory predictio n. However, expensive time consumption prevents diffusion models from real-time prediction, since a large number of denoising steps are required to assure suffi cient representation ability. To resolve the dilemma, we present LEapfrog Diffus ion model (LED), a novel diffusion-based trajectory prediction model, which prov ides real-time, precise, and diverse predictions. The core of the proposed LED i s to leverage a trainable leapfrog initializer to directly learn an expressive m ulti-modal distribution of future trajectories, which skips a large number of de noising steps, significantly accelerating inference speed. Moreover, the leapfro g initializer is trained to appropriately allocate correlated samples to provide a diversity of predicted future trajectories, significantly improving predictio n performances. Extensive experiments on four real-world datasets, including NBA /NFL/SDD/ETH-UCY, show that LED consistently improves performance and achieves 2 3.7%/21.9% ADE/FDE improvement on NFL. The proposed LED also speeds up the infer ence 19.3/30.8/24.3/25.1 times compared to the standard diffusion model on NBA/N FL/SDD/ETH-UCY, satisfying real-time inference needs. Code is available at https ://github.com/MediaBrain-SJTU/LED.
****
DaFKD: Domain-Aware Federated Knowledge Distillation
Haozhao Wang, Yichen Li, Wenchao Xu, Ruixuan Li, Yufeng Zhan, Zhigang Zeng; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2023, pp. 20412-20421
Federated Distillation (FD) has recently attracted increasing attention for its efficiency in aggregating multiple diverse local models trained from statistical ly heterogeneous data of distributed clients. Existing FD methods generally trea t these models equally by merely computing the average of their output soft pred ictions for some given input distillation sample, which does not take the divers ity across all local models into account, thus leading to degraded performance o f the aggregated model, especially when some local models learn little knowledge about the sample. In this paper, we propose a new perspective that treats the l ocal data in each client as a specific domain and design a novel domain knowledg e aware federated distillation method, dubbed DaFKD, that can discern the import ance of each model to the distillation sample, and thus is able to optimize the ensemble of soft predictions from diverse models. Specifically, we employ a doma in discriminator for each client, which is trained to identify the correlation f actor between the sample and the corresponding domain. Then, to facilitate the t raining of the domain discriminator while saving communication costs, we propose sharing its partial parameters with the classification model. Extensive experim ents on various datasets and settings show that the proposed method can improve the model accuracy by up to 6.02% compared to state-of-the-art baselines.
****

GeoLayoutLM: Geometric Pre-Training for Visual Information Extraction

Chuwei Luo, Changxu Cheng, Qi Zheng, Cong Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7092-7101

Visual information extraction (VIE) plays an important role in Document Intelligence. Generally, it is divided into two tasks: semantic entity recognition (SER) and relation extraction (RE). Recently, pre-trained models for documents have achieved substantial progress in VIE, particularly in SER. However, most of the existing models learn the geometric representation in an implicit way, which has been found insufficient for the RE task since geometric information is especially crucial for RE. Moreover, we reveal another factor that limits the performance of RE lies in the objective gap between the pre-training phase and the fine-tuning phase for RE. To tackle these issues, we propose in this paper a multi-modal framework, named GeoLayoutLM, for VIE. GeoLayoutLM explicitly models the geometric relations in pre-training, which we call geometric pre-training. Geometric pre-training is achieved by three specially designed geometry-related pre-training tasks. Additionally, novel relation heads, which are pre-trained by the geometric pre-training tasks and fine-tuned for RE, are elaborately designed to enrich and enhance the feature representation. According to extensive experiments on standard VIE benchmarks, GeoLayoutLM achieves highly competitive scores in the SER task and significantly outperforms the previous state-of-the-arts for RE (e.g.,the F1 score of RE on FUNSD is boosted from 80.35% to 89.45%).

****

Class-Incremental Exemplar Compression for Class-Incremental Learning

Zilin Luo, Yaoyao Liu, Bernt Schiele, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11371-11380

Exemplar-based class-incremental learning (CIL) finetunes the model with all samples of new classes but few-shot exemplars of old classes in each incremental phase, where the "few-shot" abides by the limited memory budget. In this paper, we break this "few-shot" limit based on a simple yet surprisingly effective idea: compressing exemplars by downsampling non-discriminative pixels and saving "many-shot" compressed exemplars in the memory. Without needing any manual annotation, we achieve this compression by generating 0-1 masks on discriminative pixels from class activation maps (CAM). We propose an adaptive mask generation model called class-incremental masking (CIM) to explicitly resolve two difficulties of using CAM: 1) transforming the heatmaps of CAM to 0-1 masks with an arbitrary threshold leads to a trade-off between the coverage on discriminative pixels and the quantity of exemplars, as the total memory is fixed; and 2) optimal thresholds vary for different object classes, which is particularly obvious in the dynamic environment of CIL. We optimize the CIM model alternatively with the conventional CIL model through a bilevel optimization problem. We conduct extensive experiments on high-resolution CIL benchmarks including Food-101, ImageNet-100, and ImageNet-1000, and show that using the compressed exemplars by CIM can achieve a new state-of-the-art CIL accuracy, e.g., 4.8 percentage points higher than FOSTER on 10-Phase ImageNet-1000. Our code is available at https://github.com/xfflzl/CIM-CIL.

****

Boost Vision Transformer With GPU-Friendly Sparsity and Quantization

Chong Yu, Tao Chen, Zhongxue Gan, Jiayuan Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22658-22668

The transformer extends its success from the language to the vision domain. Because of the numerous stacked self-attention and cross-attention blocks in the transformer, which involve many high-dimensional tensor multiplication operations, the acceleration deployment of vision transformer on GPU hardware is challenging and also rarely studied. This paper thoroughly designs a compression scheme to maximally utilize the GPU-friendly 2:4 fine-grained structured sparsity and quantization. Specially, an original large model with dense weight parameters is first pruned into a sparse one by 2:4 structured pruning, which considers the GPU's acceleration of 2:4 structured sparse pattern with FP16 data type, then the floating-point sparse model is further quantized into a fixed-point one by sparse-d

istillation-aware quantization aware training, which considers GPU can provide an extra speedup of 2:4 sparse calculation with integer tensors. A mixed-strategy knowledge distillation is used during the pruning and quantization process. The proposed compression scheme is flexible to support supervised and unsupervised learning styles. Experiment results show GPUSQ-ViT scheme achieves state-of-the-art compression by reducing vision transformer models 6.4-12.7 times on model size and 30.3-62 times on FLOPs with negligible accuracy degradation on ImageNet classification, COCO detection and ADE20K segmentation benchmarking tasks. Moreover, GPUSQ-ViT can boost actual deployment performance by 1.39-1.79 times and 3.22-3.43 times of latency and throughput on A100 GPU, and 1.57-1.69 times and 2.11-2.51 times improvement of latency and throughput on AGX Orin.
****
Spectral Bayesian Uncertainty for Image Super-Resolution
Tao Liu, Jun Cheng, Shan Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18166-18175
Recently deep learning techniques have significantly advanced image super-resolution (SR). Due to the black-box nature, quantifying reconstruction uncertainty is crucial when employing these deep SR networks. Previous approaches for SR uncertainty estimation mostly focus on capturing pixel-wise uncertainty in the spatial domain. SR uncertainty in the frequency domain which is highly related to image SR is seldom explored. In this paper, we propose to quantify spectral Bayesian uncertainty in image SR. To achieve this, a Dual-Domain Learning (DDL) framework is first proposed. Combined with Bayesian approaches, the DDL model is able to estimate spectral uncertainty accurately, enabling a reliability assessment for high frequencies reasoning from the frequency domain perspective. Extensive experiments under non-ideal premises are conducted and demonstrate the effectiveness of the proposed spectral uncertainty. Furthermore, we propose a novel Spectral Uncertainty based Decoupled Frequency (SUDF) training scheme for perceptual SR. Experimental results show the proposed SUDF can evidently boost perceptual quality of SR results without sacrificing much pixel accuracy.
****
Behind the Scenes: Density Fields for Single View Reconstruction
Felix Wimbauer, Nan Yang, Christian Rupprecht, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9076-9086
Inferring a meaningful geometric scene representation from a single image is a fundamental problem in computer vision. Approaches based on traditional depth map prediction can only reason about areas that are visible in the image. Currently, neural radiance fields (NeRFs) can capture true 3D including color, but are too complex to be generated from a single image. As an alternative, we propose to predict an implicit density field from a single image. It maps every location in the frustum of the image to volumetric density. By directly sampling color from the available views instead of storing color in the density field, our scene representation becomes significantly less complex compared to NeRFs, and a neural network can predict it in a single forward pass. The network is trained through self-supervision from only video data. Our formulation allows volume rendering to perform both depth prediction and novel view synthesis. Through experiments, we show that our method is able to predict meaningful geometry for regions that are occluded in the input image. Additionally, we demonstrate the potential of our approach on three datasets for depth prediction and novel-view synthesis.
****
StyleGAN Salon: Multi-View Latent Optimization for Pose-Invariant Hairstyle Transfer
Sasikarn Khwanmuang, Pakkapon Phongthawee, Patsorn Sangkloy, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8609-8618
Our paper seeks to transfer the hairstyle of a reference image to an input photo for virtual hair try-on. We target a variety of challenges scenarios, such as transforming a long hairstyle with bangs to a pixie cut, which requires removing the existing hair and inferring how the forehead would look, or transferring par

tially visible hair from a hat-wearing person in a different pose. Past solution s leverage StyleGAN for hallucinating any missing parts and producing a seamless face-hair composite through so-called GAN inversion or projection. However, the re remains a challenge in controlling the hallucinations to accurately transfer hairstyle and preserve the face shape and identity of the input. To overcome thi s, we propose a multi-view optimization framework that uses "two different views " of reference composites to semantically guide occluded or ambiguous regions. O ur optimization shares information between two poses, which allows us to produce high fidelity and realistic results from incomplete references. Our framework p roduces high-quality results and outperforms prior work in a user study that con sists of significantly more challenging hair transfer scenarios than previously studied. Project page: https://stylegan-salon.github.io/.
****
Resource-Efficient RGBD Aerial Tracking
Jinyu Yang, Shang Gao, Zhe Li, Feng Zheng, Aleš Leonardis; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1 3374-13383
Aerial robots are now able to fly in complex environments, and drone-captured da ta gains lots of attention in object tracking. However, current research on aeri al perception has mainly focused on limited categories, such as pedestrian or ve hicle, and most scenes are captured in urban environments from a birds-eye view. Recently, UAVs equipped with depth cameras have been also deployed for more com plex applications, while RGBD aerial tracking is still unexplored. Compared with traditional RGB object tracking, adding depth information can more effectively deal with more challenging scenes such as target and background interference. To this end, in this paper, we explore RGBD aerial tracking in an overhead space, which can greatly enlarge the development of drone-based visual perception. To b oost the research, we first propose a large-scale benchmark for RGBD aerial trac king, containing 1,000 drone-captured RGBD videos with dense annotations. Then, as drone-based applications require for real-time processing with limited comput ational resources, we also propose an efficient RGBD tracker named EMT. Our trac ker runs at over 100 fps on GPU, and 25 fps on the edge platform of NVidia Jetso n NX Xavier, benefiting from its efficient multimodal fusion and feature matchin g. Extensive experiments show that our EMT achieves promising tracking performan ce. All resources are available at https://github.com/yjybuaa/RGBDAerialTracking .
****
Mutual Information-Based Temporal Difference Learning for Human Pose Estimation in Video
Runyang Feng, Yixing Gao, Xueqing Ma, Tze Ho Elden Tse, Hyung Jin Chang; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17131-17141
Temporal modeling is crucial for multi-frame human pose estimation. Most existin g methods directly employ optical flow or deformable convolution to predict full -spectrum motion fields, which might incur numerous irrelevant cues, such as a n earby person or background. Without further efforts to excavate meaningful motio n priors, their results are suboptimal, especially in complicated spatiotemporal interactions. On the other hand, the temporal difference has the ability to enc ode representative motion information which can potentially be valuable for pose estimation but has not been fully exploited. In this paper, we present a novel multi-frame human pose estimation framework, which employs temporal differences across frames to model dynamic contexts and engages mutual information objective ly to facilitate useful motion information disentanglement. To be specific, we d esign a multi-stage Temporal Difference Encoder that performs incremental cascad ed learning conditioned on multi-stage feature difference sequences to derive in formative motion representation. We further propose a Representation Disentangle ment module from the mutual information perspective, which can grasp discriminat ive task-relevant motion signals by explicitly defining useful and noisy constit uents of the raw motion features and minimizing their mutual information. These place us to rank No.1 in the Crowd Pose Estimation in Complex Events Challenge o

n benchmark dataset HiEve, and achieve state-of-the-art performance on three ben chmarks PoseTrack2017, PoseTrack2018, and PoseTrack21.
****
Bilateral Memory Consolidation for Continual Learning
Xing Nie, Shixiong Xu, Xiyan Liu, Gaofeng Meng, Chunlei Huo, Shiming Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16026-16035
Humans are proficient at continuously acquiring and integrating new knowledge. By contrast, deep models forget catastrophically, especially when tackling highly long task sequences. Inspired by the way our brains constantly rewrite and consolidate past recollections, we propose a novel Bilateral Memory Consolidation (BiMeCo) framework that focuses on enhancing memory interaction capabilities. Specifically, BiMeCo explicitly decouples model parameters into short-term memory module and long-term memory module, responsible for representation ability of the model and generalization over all learned tasks, respectively. BiMeCo encourages dynamic interactions between two memory modules by knowledge distillation and momentum-based updating for forming generic knowledge to prevent forgetting. The proposed BiMeCo is parameter-efficient and can be integrated into existing methods seamlessly. Extensive experiments on challenging benchmarks show that BiMeCo significantly improves the performance of existing continual learning methods. For example, combined with the state-of-the-art method CwD, BiMeCo brings in significant gains of around 2% to 6% while using 2x fewer parameters on CIFAR-100 under ResNet-18.
****
SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision
Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, Maja Pantic, Christian Fuegen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18806-18815
Recently reported state-of-the-art results in visual speech recognition (VSR) often rely on increasingly large amounts of video data, while the publicly available transcribed video datasets are limited in size. In this paper, for the first time, we study the potential of leveraging synthetic visual data for VSR. Our method, termed SynthVSR, substantially improves the performance of VSR systems with synthetic lip movements. The key idea behind SynthVSR is to leverage a speech-driven lip animation model that generates lip movements conditioned on the input speech. The speech-driven lip animation model is trained on an unlabeled audio-visual dataset and could be further optimized towards a pre-trained VSR model when labeled videos are available. As plenty of transcribed acoustic data and face images are available, we are able to generate large-scale synthetic data using the proposed lip animation model for semi-supervised VSR training. We evaluate the performance of our approach on the largest public VSR benchmark - Lip Reading Sentences 3 (LRS3). SynthVSR achieves a WER of 43.3% with only 30 hours of real labeled data, outperforming off-the-shelf approaches using thousands of hours of video. The WER is further reduced to 27.9% when using all 438 hours of labeled data from LRS3, which is on par with the state-of-the-art self-supervised AV-HuBERT method. Furthermore, when combined with large-scale pseudo-labeled audio-visual data SynthVSR yields a new state-of-the-art VSR WER of 16.9% using publicly available data only, surpassing the recent state-of-the-art approaches trained with 29 times more non-public machine-transcribed video data (90,000 hours). Finally, we perform extensive ablation studies to understand the effect of each component in our proposed method.
****
BiasBed - Rigorous Texture Bias Evaluation
Nikolai Kalischek, Rodrigo Caye Daudt, Torben Peters, Reinhard Furrer, Jan D. Wegner, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22221-22230
The well-documented presence of texture bias in modern convolutional neural networks has led to a plethora of algorithms that promote an emphasis on shape cues, often to support generalization to new domains. Yet, common datasets, benchmark

s and general model selection strategies are missing, and there is no agreed, rigorous evaluation protocol. In this paper, we investigate difficulties and limitations when training networks with reduced texture bias. In particular, we also show that proper evaluation and meaningful comparisons between methods are not trivial. We introduce BiasBed, a testbed for texture- and style-biased training, including multiple datasets and a range of existing algorithms. It comes with an extensive evaluation protocol that includes rigorous hypothesis testing to gauge the significance of the results, despite the considerable training instability of some style bias methods. Our extensive experiments, shed new light on the need for careful, statistically founded evaluation protocols for style bias (and beyond). E.g., we find that some algorithms proposed in the literature do not significantly mitigate the impact of style bias at all. With the release of BiasBed, we hope to foster a common understanding of consistent and meaningful comparisons, and consequently faster progress towards learning methods free of texture bias. Code is available at https://github.com/D1noFuzi/BiasBed.
****
Open-Category Human-Object Interaction Pre-Training via Language Modeling Framework
Sipeng Zheng, Boshen Xu, Qin Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19392-19402
Human-object interaction (HOI) has long been plagued by the conflict between limited supervised data and a vast number of possible interaction combinations in real life. Current methods trained from closed-set data predict HOIs as fixed-dimension logits, which restricts their scalability to open-set categories. To address this issue, we introduce OpenCat, a language modeling framework that reformulates HOI prediction as sequence generation. By converting HOI triplets into a token sequence through a serialization scheme, our model is able to exploit the open-set vocabulary of the language modeling framework to predict novel interaction classes with a high degree of freedom. In addition, inspired by the great success of vision-language pre-training, we collect a large amount of weakly-supervised data related to HOI from image-caption pairs, and devise several auxiliary proxy tasks, including soft relational matching and human-object relation prediction, to pre-train our model. Extensive experiments show that our OpenCat significantly boosts HOI performance, particularly on a broad range of rare and unseen categories.
****
SFD2: Semantic-Guided Feature Detection and Description
Fei Xue, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5206-5216
Visual localization is a fundamental task for various applications including autonomous driving and robotics. Prior methods focus on extracting large amounts of often redundant locally reliable features, resulting in limited efficiency and accuracy, especially in large-scale environments under challenging conditions. Instead, we propose to extract globally reliable features by implicitly embedding high-level semantics into both the detection and description processes. Specifically, our semantic-aware detector is able to detect keypoints from reliable regions (e.g. building, traffic lane) and suppress reliable areas (e.g. sky, car) implicitly instead of relying on explicit semantic labels. This boosts the accuracy of keypoint matching by reducing the number of features sensitive to appearance changes and avoiding the need of additional segmentation networks at test time. Moreover, our descriptors are augmented with semantics and have stronger discriminative ability, providing more inliers at test time. Particularly, experiments on long-term large-scale visual localization Aachen Day-Night and RobotCar-Seasons datasets demonstrate that our model outperforms previous local features and gives competitive accuracy to advanced matchers but is about 2 and 3 times faster when using 2k and 4k keypoints, respectively.
****
Search-Map-Search: A Frame Selection Paradigm for Action Recognition
Mingjun Zhao, Yakun Yu, Xiaoli Wang, Lei Yang, Di Niu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10627

-10636
Despite the success of deep learning in video understanding tasks, processing every frame in a video is computationally expensive and often unnecessary in real-time applications. Frame selection aims to extract the most informative and representative frames to help a model better understand video content. Existing frame selection methods either individually sample frames based on per-frame importance prediction, without considering interaction among frames, or adopt reinforcement learning agents to find representative frames in succession, which are costly to train and may lead to potential stability issues. To overcome the limitations of existing methods, we propose a Search-Map-Search learning paradigm which combines the advantages of heuristic search and supervised learning to select the best combination of frames from a video as one entity. By combining search with learning, the proposed method can better capture frame interactions while incurring a low inference overhead. Specifically, we first propose a hierarchical search method conducted on each training video to search for the optimal combination of frames with the lowest error on the downstream task. A feature mapping function is then learned to map the frames of a video to the representation of its target optimal frame combination. During inference, another search is performed on an unseen video to select a combination of frames whose feature representation is close to the projected feature representation. Extensive experiments based on several action recognition benchmarks demonstrate that our frame selection method effectively improves performance of action recognition models, and significantly outperforms a number of competitive baselines.
****
Uncovering the Missing Pattern: Unified Framework Towards Trajectory Imputation and Prediction
Yi Xu, Armin Bazarjani, Hyung-gun Chi, Chiho Choi, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9632-9643
Trajectory prediction is a crucial undertaking in understanding entity movement or human behavior from observed sequences. However, current methods often assume that the observed sequences are complete while ignoring the potential for missing values caused by object occlusion, scope limitation, sensor failure, etc. This limitation inevitably hinders the accuracy of trajectory prediction. To address this issue, our paper presents a unified framework, the Graph-based Conditional Variational Recurrent Neural Network (GC-VRNN), which can perform trajectory imputation and prediction simultaneously. Specifically, we introduce a novel Multi-Space Graph Neural Network (MS-GNN) that can extract spatial features from incomplete observations and leverage missing patterns. Additionally, we employ a Conditional VRNN with a specifically designed Temporal Decay (TD) module to capture temporal dependencies and temporal missing patterns in incomplete trajectories. The inclusion of the TD module allows for valuable information to be conveyed through the temporal flow. We also curate and benchmark three practical datasets for the joint problem of trajectory imputation and prediction. Extensive experiments verify the exceptional performance of our proposed method. As far as we know, this is the first work to address the lack of benchmarks and techniques for trajectory imputation and prediction in a unified manner.
****
CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not
Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2765-2775
In this paper, we leverage CLIP for zero-shot sketch based image retrieval (ZS-SBIR). We are largely inspired by recent advances on foundation models and the unparalleled generalisation ability they seem to offer, but for the first time tailor it to benefit the sketch community. We put forward novel designs on how best to achieve this synergy, for both the category setting and the fine-grained setting ("all"). At the very core of our solution is a prompt learning setup. First we show just via factoring in sketch-specific prompts, we already have a category-level ZS-SBIR system that overshoots all prior arts, by a large margin (24.8%

) - a great testimony on studying the CLIP and ZS-SBIR synergy. Moving onto the
fine-grained setup is however trickier, and requires a deeper dive into this syn
ergy. For that, we come up with two specific designs to tackle the fine-grained
matching nature of the problem: (i) an additional regularisation loss to ensure
the relative separation between sketches and photos is uniform across categories
, which is not the case for the gold standard standalone triplet loss, and (ii)
a clever patch shuffling technique to help establishing instance-level structura
l correspondences between sketch-photo pairs. With these designs, we again obser
ve significant performance gains in the region of 26.9% over previous state-of-t
he-art. The take-home message, if any, is the proposed CLIP and prompt learning
paradigm carries great promise in tackling other sketch-related tasks (not limit
ed to ZS-SBIR) where data scarcity remains a great challenge. Project page: http
s://aneeshan95.github.io/Sketch_LVM/
****
FlexiViT: One Model for All Patch Sizes
Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornbli
th, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, F
ilip Pavetic; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 14496-14506
Vision Transformers convert images to sequences by slicing them into patches. Th
e size of these patches controls a speed/accuracy tradeoff, with smaller patches
 leading to higher accuracy at greater computational cost, but changing the patc
h size typically requires retraining the model. In this paper, we demonstrate th
at simply randomizing the patch size at training time leads to a single set of w
eights that performs well across a wide range of patch sizes, making it possible
 to tailor the model to different compute budgets at deployment time. We extensi
vely evaluate the resulting model, which we call FlexiViT, on a wide range of ta
sks, including classification, image-text retrieval, openworld detection, panopt
ic segmentation, and semantic segmentation, concluding that it usually matches,
and sometimes outperforms, standard ViT models trained at a single patch size in
 an otherwise identical setup. Hence, FlexiViT training is a simple drop-in impr
ovement for ViT that makes it easy to add compute-adaptive capabilities to most
models relying on a ViT backbone architecture. Code and pretrained models are av
ailable at github.com/googleresearch/big_vision.
****
RIAV-MVS: Recurrent-Indexing an Asymmetric Volume for Multi-View Stereo
Changjiang Cai, Pan Ji, Qingan Yan, Yi Xu; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 919-928
This paper presents a learning-based method for multi-view depth estimation from
 posed images. Our core idea is a "learning-to-optimize" paradigm that iterative
ly indexes a plane-sweeping cost volume and regresses the depth map via a convol
utional Gated Recurrent Unit (GRU). Since the cost volume plays a paramount role
 in encoding the multi-view geometry, we aim to improve its construction both at
 pixel- and frame- levels. At the pixel level, we propose to break the symmetry
of the Siamese network (which is typically used in MVS to extract image features
) by introducing a transformer block to the reference image (but not to the sour
ce images). Such an asymmetric volume allows the network to extract global featu
res from the reference image to predict its depth map. Given potential inaccurac
ies in the poses between reference and source images, we propose to incorporate
a residual pose network to correct the relative poses. This essentially rectifie
s the cost volume at the frame level. We conduct extensive experiments on real-w
orld MVS datasets and show that our method achieves state-of-the-art performance
 in terms of both within-dataset evaluation and cross-dataset generalization.
****
Structured Kernel Estimation for Photon-Limited Deconvolution
Yash Sanghvi, Zhiyuan Mao, Stanley H. Chan; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9863-9872
Images taken in a low light condition with the presence of camera shake suffer f
rom motion blur and photon shot noise. While state-of-the-art image restoration
networks show promising results, they are largely limited to well-illuminated sc

enes and their performance drops significantly when photon shot noise is strong. In this paper, we propose a new blur estimation technique customized for photon-limited conditions. The proposed method employs a gradient-based backpropagation method to estimate the blur kernel. By modeling the blur kernel using a low-dimensional representation with the key points on the motion trajectory, we significantly reduce the search space and improve the regularity of the kernel estimation problem. When plugged into an iterative framework, our novel low-dimensional representation provides improved kernel estimates and hence significantly better deconvolution performance when compared to end-to-end trained neural networks.
****
Explicit Boundary Guided Semi-Push-Pull Contrastive Learning for Supervised Anomaly Detection
Xincheng Yao, Ruoqi Li, Jing Zhang, Jun Sun, Chongyang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24490-24499
Most anomaly detection (AD) models are learned using only normal samples in an unsupervised way, which may result in ambiguous decision boundary and insufficient discriminability. In fact, a few anomaly samples are often available in real-world applications, the valuable knowledge of known anomalies should also be effectively exploited. However, utilizing a few known anomalies during training may cause another issue that the model may be biased by those known anomalies and fail to generalize to unseen anomalies. In this paper, we tackle supervised anomaly detection, i.e., we learn AD models using a few available anomalies with the objective to detect both the seen and unseen anomalies. We propose a novel explicit boundary guided semi-push-pull contrastive learning mechanism, which can enhance model's discriminability while mitigating the bias issue. Our approach is based on two core designs: First, we find an explicit and compact separating boundary as the guidance for further feature learning. As the boundary only relies on the normal feature distribution, the bias problem caused by a few known anomalies can be alleviated. Second, a boundary guided semi-push-pull loss is developed to only pull the normal features together while pushing the abnormal features apart from the separating boundary beyond a certain margin region. In this way, our model can form a more explicit and discriminative decision boundary to distinguish known and also unseen anomalies from normal samples more effectively. Code will be available at https://github.com/xcyao00/BGAD.
****
3D Video Loops From Asynchronous Input
Li Ma, Xiaoyu Li, Jing Liao, Pedro V. Sander; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 310-320
Looping videos are short video clips that can be looped endlessly without visible seams or artifacts. They provide a very attractive way to capture the dynamism of natural scenes. Existing methods have been mostly limited to 2D representations. In this paper, we take a step forward and propose a practical solution that enables an immersive experience on dynamic 3D looping scenes. The key challenge is to consider the per-view looping conditions from asynchronous input while maintaining view consistency for the 3D representation. We propose a novel sparse 3D video representation, namely Multi-Tile Video (MTV), which not only provides a view-consistent prior, but also greatly reduces memory usage, making the optimization of a 4D volume tractable. Then, we introduce a two-stage pipeline to construct the 3D looping MTV from completely asynchronous multi-view videos with no time overlap. A novel looping loss based on video temporal retargeting algorithms is adopted during the optimization to loop the 3D scene. Experiments of our framework have shown promise in successfully generating and rendering photorealistic 3D looping videos in real time even on mobile devices. The code, dataset, and live demos are available in https://limacv.github.io/VideoLoop3D_web/.
****
Style Projected Clustering for Domain Generalized Semantic Segmentation
Wei Huang, Chang Chen, Yong Li, Jiacheng Li, Cheng Li, Fenglong Song, Youliang Yan, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3061-3071

Existing semantic segmentation methods improve generalization capability, by regularizing various images to a canonical feature space. While this process contributes to generalization, it weakens the representation inevitably. In contrast to existing methods, we instead utilize the difference between images to build a better representation space, where the distinct style features are extracted and stored as the bases of representation. Then, the generalization to unseen image styles is achieved by projecting features to this known space. Specifically, we realize the style projection as a weighted combination of stored bases, where the similarity distances are adopted as the weighting factors. Based on the same concept, we extend this process to the decision part of model and promote the generalization of semantic prediction. By measuring the similarity distances to semantic bases (i.e., prototypes), we replace the common deterministic prediction with semantic clustering. Comprehensive experiments demonstrate the advantage of proposed method to the state of the art, up to 3.6% mIoU improvement in average on unseen scenarios.

****

DIP: Dual Incongruity Perceiving Network for Sarcasm Detection

Changsong Wen, Guoli Jia, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2540-2550

Sarcasm indicates the literal meaning is contrary to the real attitude. Considering the popularity and complementarity of image-text data, we investigate the task of multi-modal sarcasm detection. Different from other multi-modal tasks, for the sarcastic data, there exists intrinsic incongruity between a pair of image and text as demonstrated in psychological theories. To tackle this issue, we propose a Dual Incongruity Perceiving (DIP) network consisting of two branches to mine the sarcastic information from factual and affective levels. For the factual aspect, we introduce a channel-wise reweighting strategy to obtain semantically discriminative embeddings, and leverage gaussian distribution to model the uncertain correlation caused by the incongruity. The distribution is generated from the latest data stored in the memory bank, which can adaptively model the difference of semantic similarity between sarcastic and non-sarcastic data. For the affective aspect, we utilize siamese layers with shared parameters to learn cross-modal sentiment information. Furthermore, we use the polarity value to construct a relation graph for the mini-batch, which forms the continuous contrastive loss to acquire affective embeddings. Extensive experiments demonstrate that our proposed method performs favorably against state-of-the-art approaches. Our code is released on https://github.com/downdric/MSD.

****

Frame Interpolation Transformer and Uncertainty Guidance

Markus Plack, Karlis Martins Briedis, Abdelaziz Djelouah, Matthias B. Hullin, Markus Gross, Christopher Schroers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9811-9821

Video frame interpolation has seen important progress in recent years, thanks to developments in several directions. Some works leverage better optical flow methods with improved splatting strategies or additional cues from depth, while others have investigated alternative approaches through direct predictions or transformers. Still, the problem remains unsolved in more challenging conditions such as complex lighting or large motion. In this work, we are bridging the gap towards video production with a novel transformer-based interpolation network architecture capable of estimating the expected error together with the interpolated frame. This offers several advantages that are of key importance for frame interpolation usage: First, we obtained improved visual quality over several datasets. The improvement in terms of quality is also clearly demonstrated through a user study. Second, our method estimates error maps for the interpolated frame, which are essential for real-life applications on longer video sequences where problematic frames need to be flagged. Finally, for rendered content a partial rendering pass of the intermediate frame, guided by the predicted error, can be utilized during the interpolation to generate a new frame of superior quality. Through this error estimation, our method can produce even higher-quality intermediate frames using only a fraction of the time compared to a full rendering.

****
Learning To Generate Language-Supervised and Open-Vocabulary Scene Graph Using Pre-Trained Visual-Semantic Space
Yong Zhang, Yingwei Pan, Ting Yao, Rui Huang, Tao Mei, Chang-Wen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2915-2924
Scene graph generation (SGG) aims to abstract an image into a graph structure, by representing objects as graph nodes and their relations as labeled edges. However, two knotty obstacles limit the practicability of current SGG methods in real-world scenarios: 1) training SGG models requires time-consuming ground-truth annotations, and 2) the closed-set object categories make the SGG models limited in their ability to recognize novel objects outside of training corpora. To address these issues, we novelly exploit a powerful pre-trained visual-semantic space (VSS) to trigger language-supervised and open-vocabulary SGG in a simple yet effective manner. Specifically, cheap scene graph supervision data can be easily obtained by parsing image language descriptions into semantic graphs. Next, the noun phrases on such semantic graphs are directly grounded over image regions through region-word alignment in the pre-trained VSS. In this way, we enable open-vocabulary object detection by performing object category name grounding with a text prompt in this VSS. On the basis of visually-grounded objects, the relation representations are naturally built for relation recognition, pursuing open-vocabulary SGG. We validate our proposed approach with extensive experiments on the Visual Genome benchmark across various SGG scenarios (i.e., supervised / language-supervised, closed-set / open-vocabulary). Consistent superior performances are achieved compared with existing methods, demonstrating the potential of exploiting pre-trained VSS for SGG in more practical scenarios.
****
VectorFloorSeg: Two-Stream Graph Attention Network for Vectorized Roughcast Floorplan Segmentation
Bingchen Yang, Haiyong Jiang, Hao Pan, Jun Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1358-1367
Vector graphics (VG) are ubiquitous in industrial designs. In this paper, we address semantic segmentation of a typical VG, i.e., roughcast floorplans with bare wall structures, whose output can be directly used for further applications like interior furnishing and room space modeling. Previous semantic segmentation works mostly process well-decorated floorplans in raster images and usually yield aliased boundaries and outlier fragments in segmented rooms, due to pixel-level segmentation that ignores the regular elements (e.g. line segments) in vector floorplans. To overcome these issues, we propose to fully utilize the regular elements in vector floorplans for more integral segmentation. Our pipeline predicts room segmentation from vector floorplans by dually classifying line segments as room boundaries, and regions partitioned by line segments as room segments. To fully exploit the structural relationships between lines and regions, we use two-stream graph neural networks to process the line segments and partitioned regions respectively, and devise a novel modulated graph attention layer to fuse the heterogeneous information from one stream to the other. Extensive experiments show that by directly operating on vector floorplans, we outperform image-based methods in both mIoU and mAcc. In addition, we propose a new metric that captures room integrity and boundary regularity, which confirms that our method produces much more regular segmentations. Source code is available at https://github.com/DrZiji/VecFloorSeg
****
Neural Preset for Color Style Transfer
Zhanghan Ke, Yuhao Liu, Lei Zhu, Nanxuan Zhao, Rynson W.H. Lau; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14173-14182
In this paper, we present a Neural Preset technique to address the limitations of existing color style transfer methods, including visual artifacts, vast memory requirement, and slow style switching speed. Our method is based on two core designs. First, we propose Deterministic Neural Color Mapping (DNCM) to consistent

ly operate on each pixel via an image-adaptive color mapping matrix, avoiding artifacts and supporting high-resolution inputs with a small memory footprint. Second, we develop a two-stage pipeline by dividing the task into color normalization and stylization, which allows efficient style switching by extracting color styles as presets and reusing them on normalized input images. Due to the unavailability of pairwise datasets, we describe how to train Neural Preset via a self-supervised strategy. Various advantages of Neural Preset over existing methods are demonstrated through comprehensive evaluations. Besides, we show that our trained model can naturally support multiple applications without fine-tuning, including low-light image enhancement, underwater image correction, image dehazing, and image harmonization.
****
DeCo: Decomposition and Reconstruction for Compositional Temporal Grounding via Coarse-To-Fine Contrastive Ranking
Lijin Yang, Quan Kong, Hsuan-Kung Yang, Wadim Kehl, Yoichi Sato, Norimasa Kobori; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23130-23140
Understanding dense action in videos is a fundamental challenge towards the generalization of vision models. Several works show that compositionality is key to achieving generalization by combining known primitive elements, especially for handling novel composited structures. Compositional temporal grounding is the task of localizing dense action by using known words combined in novel ways in the form of novel query sentences for the actual grounding. In recent works, composition is assumed to be learned from pairs of whole videos and language embeddings through large scale self-supervised pre-training. Alternatively, one can process the video and language into word-level primitive elements, and then only learn fine-grained semantic correspondences. Both approaches do not consider the granularity of the compositions, where different query granularity corresponds to different video segments. Therefore, a good compositional representation should be sensitive to different video and query granularity. We propose a method to learn a coarse-to-fine compositional representation by decomposing the original query sentence into different granular levels, and then learning the correct correspondences between the video and recombined queries through a contrastive ranking constraint. Additionally, we run temporal boundary prediction in a coarse-to-fine manner for precise grounding boundary detection. Experiments are performed on two datasets Charades-CG and ActivityNet-CG showing the superior compositional generalizability of our approach.
****
Dynamic Aggregated Network for Gait Recognition
Kang Ma, Ying Fu, Dezhi Zheng, Chunshui Cao, Xuecai Hu, Yongzhen Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22076-22085
Gait recognition is beneficial for a variety of applications, including video surveillance, crime scene investigation, and social security, to mention a few. However, gait recognition often suffers from multiple exterior factors in real scenes, such as carrying conditions, wearing overcoats, and diverse viewing angles. Recently, various deep learning-based gait recognition methods have achieved promising results, but they tend to extract one of the salient features using fixed-weighted convolutional networks, do not well consider the relationship within gait features in key regions, and ignore the aggregation of complete motion patterns. In this paper, we propose a new perspective that actual gait features include global motion patterns in multiple key regions, and each global motion pattern is composed of a series of local motion patterns. To this end, we propose a Dynamic Aggregation Network (DANet) to learn more discriminative gait features. Specifically, we create a dynamic attention mechanism between the features of neighboring pixels that not only adaptively focuses on key regions but also generates more expressive local motion patterns. In addition, we develop a self-attention mechanism to select representative local motion patterns and further learn robust global motion patterns. Extensive experiments on three popular public gait datasets, i.e., CASIA-B, OUMVLP, and Gait3D, demonstrate that the proposed metho

d can provide substantial improvements over the current state-of-the-art methods
.
****
Wavelet Diffusion Models Are Fast and Scalable Image Generators
Hao Phung, Quan Dao, Anh Tran; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2023, pp. 10199-10208
Diffusion models are rising as a powerful solution for high-fidelity image gener
ation, which exceeds GANs in quality in many circumstances. However, their slow
training and inference speed is a huge bottleneck, blocking them from being used
 in real-time applications. A recent DiffusionGAN method significantly decreases
 the models' running time by reducing the number of sampling steps from thousand
s to several, but their speeds still largely lag behind the GAN counterparts. Th
is paper aims to reduce the speed gap by proposing a novel wavelet-based diffusi
on scheme. We extract low-and-high frequency components from both image and feat
ure levels via wavelet decomposition and adaptively handle these components for
faster processing while maintaining good generation quality. Furthermore, we pro
pose to use a reconstruction term, which effectively boosts the model training c
onvergence. Experimental results on CelebA-HQ, CIFAR-10, LSUN-Church, and STL-10
 datasets prove our solution is a stepping-stone to offering real-time and high-
fidelity diffusion models. Our code and pre-trained checkpoints are available at
 https://github.com/VinAIResearch/WaveDiff.git.
****
PA&DA: Jointly Sampling Path and Data for Consistent NAS
Shun Lu, Yu Hu, Longxing Yang, Zihao Sun, Jilin Mei, Jianchao Tan, Chengru Song;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2023, pp. 11940-11949
Based on the weight-sharing mechanism, one-shot NAS methods train a supernet and
 then inherit the pre-trained weights to evaluate sub-models, largely reducing t
he search cost. However, several works have pointed out that the shared weights
suffer from different gradient descent directions during training. And we furthe
r find that large gradient variance occurs during supernet training, which degra
des the supernet ranking consistency. To mitigate this issue, we propose to expl
icitly minimize the gradient variance of the supernet training by jointly optimi
zing the sampling distributions of PAth and DAta (PA&DA). We theoretically deriv
e the relationship between the gradient variance and the sampling distributions,
 and reveal that the optimal sampling probability is proportional to the normali
zed gradient norm of path and training data. Hence, we use the normalized gradie
nt norm as the importance indicator for path and training data, and adopt an imp
ortance sampling strategy for the supernet training. Our method only requires ne
gligible computation cost for optimizing the sampling distributions of path and
data, but achieves lower gradient variance during supernet training and better g
eneralization performance for the supernet, resulting in a more consistent NAS.
We conduct comprehensive comparisons with other improved approaches in various s
earch spaces. Results show that our method surpasses others with more reliable r
anking performance and higher accuracy of searched architectures, showing the ef
fectiveness of our method. Code is available at https://github.com/ShunLu91/PA-D
A.
****
Sphere-Guided Training of Neural Implicit Surfaces
Andreea Dogaru, Andrei-Timotei Ardelean, Savva Ignatyev, Egor Zakharov, Evgeny B
urnaev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2023, pp. 20844-20853
In recent years, neural distance functions trained via volumetric ray marching h
ave been widely adopted for multi-view 3D reconstruction. These methods, however
, apply the ray marching procedure for the entire scene volume, leading to reduc
ed sampling efficiency and, as a result, lower reconstruction quality in the are
as of high-frequency details. In this work, we address this problem via joint tr
aining of the implicit function and our new coarse sphere-based surface reconstr
uction. We use the coarse representation to efficiently exclude the empty volume
 of the scene from the volumetric ray marching procedure without additional forw

ard passes of the neural surface network, which leads to an increased fidelity o
f the reconstructions compared to the base systems. We evaluate our approach by
incorporating it into the training procedures of several implicit surface modeli
ng methods and observe uniform improvements across both synthetic and real-world
 datasets. Our codebase can be accessed via the project page.
****
3D Spatial Multimodal Knowledge Accumulation for Scene Graph Prediction in Point
 Cloud
Mingtao Feng, Haoran Hou, Liang Zhang, Zijie Wu, Yulan Guo, Ajmal Mian; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2023, pp. 9182-9191
In-depth understanding of a 3D scene not only involves locating/recognizing indi
vidual objects, but also requires to infer the relationships and interactions am
ong them. However, since 3D scenes contain partially scanned objects with physic
al connections, dense placement, changing sizes, and a wide variety of challengi
ng relationships, existing methods perform quite poorly with limited training sa
mples. In this work, we find that the inherently hierarchical structures of phys
ical space in 3D scenes aid in the automatic association of semantic and spatial
 arrangements, specifying clear patterns and leading to less ambiguous predictio
ns. Thus, they well meet the challenges due to the rich variations within scene
categories. To achieve this, we explicitly unify these structural cues of 3D phy
sical spaces into deep neural networks to facilitate scene graph prediction. Spe
cifically, we exploit an external knowledge base as a baseline to accumulate bot
h contextualized visual content and textual facts to form a 3D spatial multimoda
l knowledge graph. Moreover, we propose a knowledge-enabled scene graph predicti
on module benefiting from the 3D spatial knowledge to effectively regularize sem
antic space of relationships. Extensive experiments demonstrate the superiority
of the proposed method over current state-of-the-art competitors. Our code is av
ailable at https://github.com/HHrEtvP/SMKA.
****
Extracting Motion and Appearance via Inter-Frame Attention for Efficient Video F
rame Interpolation
Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, Limin Wang; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 5682-5692
Effectively extracting inter-frame motion and appearance information is importan
t for video frame interpolation (VFI). Previous works either extract both types
of information in a mixed way or devise separate modules for each type of inform
ation, which lead to representation ambiguity and low efficiency. In this paper,
 we propose a new module to explicitly extract motion and appearance information
 via a unified operation. Specifically, we rethink the information process in in
ter-frame attention and reuse its attention map for both appearance feature enha
ncement and motion information extraction. Furthermore, for efficient VFI, our p
roposed module could be seamlessly integrated into a hybrid CNN and Transformer
architecture. This hybrid pipeline can alleviate the computational complexity of
 inter-frame attention as well as preserve detailed low-level structure informat
ion. Experimental results demonstrate that, for both fixed- and arbitrary-timest
ep interpolation, our method achieves state-of-the-art performance on various da
tasets. Meanwhile, our approach enjoys a lighter computation overhead over model
s with close performance. The source code and models are available at https://gi
thub.com/MCG-NJU/EMA-VFI.
****
Bias Mimicking: A Simple Sampling Approach for Bias Mitigation
Maan Qraitem, Kate Saenko, Bryan A. Plummer; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20311-20320
Prior work has shown that Visual Recognition datasets frequently underrepresent
bias groups B (e.g. Female) within class labels Y (e.g. Programmers). This datas
et bias can lead to models that learn spurious correlations between class labels
 and bias groups such as age, gender, or race. Most recent methods that address
this problem require significant architectural changes or additional loss functi

ons requiring more hyper-parameter tuning. Alternatively, data sampling baselines from the class imbalance literature (eg Undersampling, Upweighting), which can often be implemented in a single line of code and often have no hyperparameters, offer a cheaper and more efficient solution. However, these methods suffer from significant shortcomings. For example, Undersampling drops a significant part of the input distribution per epoch while Oversampling repeats samples, causing overfitting. To address these shortcomings, we introduce a new class-conditioned sampling method: Bias Mimicking. The method is based on the observation that if a class c bias distribution, i.e., P_D(B|Y=c) is mimicked across every c' != c, then Y and B are statistically independent. Using this notion, BM, through a novel training procedure, ensures that the model is exposed to the entire distribution per epoch without repeating samples. Consequently, Bias Mimicking improves underrepresented groups' accuracy of sampling methods by 3% over four benchmarks while maintaining and sometimes improving performance over nonsampling methods. Code: https://github.com/mqraitem/Bias-Mimicking
****
ViTs for SITS: Vision Transformers for Satellite Image Time Series
Michail Tarasiou, Erik Chavez, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10418-10428
In this paper we introduce the Temporo-Spatial Vision Transformer (TSViT), a fully-attentional model for general Satellite Image Time Series (SITS) processing based on the Vision Transformer (ViT). TSViT splits a SITS record into non-overlapping patches in space and time which are tokenized and subsequently processed by a factorized temporo-spatial encoder. We argue, that in contrast to natural images, a temporal-then-spatial factorization is more intuitive for SITS processing and present experimental evidence for this claim. Additionally, we enhance the model's discriminative power by introducing two novel mechanisms for acquisition-time-specific temporal positional encodings and multiple learnable class tokens. The effect of all novel design choices is evaluated through an extensive ablation study. Our proposed architecture achieves state-of-the-art performance, surpassing previous approaches by a significant margin in three publicly available SITS semantic segmentation and classification datasets. All model, training and evaluation codes can be found at https://github.com/michaeltrs/DeepSatModels.
****
NoisyQuant: Noisy Bias-Enhanced Post-Training Activation Quantization for Vision Transformers
Yijiang Liu, Huanrui Yang, Zhen Dong, Kurt Keutzer, Li Du, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20321-20330
The complicated architecture and high training cost of vision transformers urge the exploration of post-training quantization. However, the heavy-tailed distribution of vision transformer activations hinders the effectiveness of previous post-training quantization methods, even with advanced quantizer designs. Instead of tuning the quantizer to better fit the complicated activation distribution, this paper proposes NoisyQuant, a quantizer-agnostic enhancement for the post-training activation quantization performance of vision transformers. We make a surprising theoretical discovery that for a given quantizer, adding a fixed Uniform noisy bias to the values being quantized can significantly reduce the quantization error under provable conditions. Building on the theoretical insight, NoisyQuant achieves the first success on actively altering the heavy-tailed activation distribution with additive noisy bias to fit a given quantizer. Extensive experiments show NoisyQuant largely improves the post-training quantization performance of vision transformer with minimal computation overhead. For instance, on linear uniform 6-bit activation quantization, NoisyQuant improves SOTA top-1 accuracy on ImageNet by up to 1.7%, 1.1% and 0.5% for ViT, DeiT, and Swin Transformer respectively, achieving on-par or even higher performance than previous nonlinear, mixed-precision quantization.
****
Semi-Supervised Stereo-Based 3D Object Detection via Cross-View Consensus

Wenhao Wu, Hau San Wong, Si Wu; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2023, pp. 17471-17481
Stereo-based 3D object detection, which aims at detecting 3D objects with stereo
 cameras, shows great potential in low-cost deployment compared to LiDAR-based m
ethods and excellent performance compared to monocular-based algorithms. However
, the impressive performance of stereo-based 3D object detection is at the huge
cost of high-quality manual annotations, which are hardly attainable for any giv
en scene. Semi-supervised learning, in which limited annotated data and numerous
 unannotated data are required to achieve a satisfactory model, is a promising m
ethod to address the problem of data deficiency. In this work, we propose to ach
ieve semi-supervised learning for stereo-based 3D object detection through pseud
o annotation generation from a temporal-aggregated teacher model, which temporal
ly accumulates knowledge from a student model. To facilitate a more stable and a
ccurate depth estimation, we introduce Temporal-Aggregation-Guided (TAG) dispari
ty consistency, a cross-view disparity consistency constraint between the teache
r model and the student model for robust and improved depth estimation. To mitig
ate noise in pseudo annotation generation, we propose a cross-view agreement str
ategy, in which pseudo annotations should attain high degree of agreements betwe
en 3D and 2D views, as well as between binocular views. We perform extensive exp
eriments on the KITTI 3D dataset to demonstrate our proposed method's capability
 in leveraging a huge amount of unannotated stereo images to attain significantl
y improved detection results.
****
Minimizing Maximum Model Discrepancy for Transferable Black-Box Targeted Attacks
Anqi Zhao, Tong Chu, Yahao Liu, Wen Li, Jingjing Li, Lixin Duan; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
 pp. 8153-8162
In this work, we study the black-box targeted attack problem from the model disc
repancy perspective. On the theoretical side, we present a generalization error
bound for black-box targeted attacks, which gives a rigorous theoretical analysi
s for guaranteeing the success of the attack. We reveal that the attack error on
 a target model mainly depends on empirical attack error on the substitute model
 and the maximum model discrepancy among substitute models. On the algorithmic s
ide, we derive a new algorithm for black-box targeted attacks based on our theor
etical analysis, in which we additionally minimize the maximum model discrepancy
(M3D) of the substitute models when training the generator to generate adversari
al examples. In this way, our model is capable of crafting highly transferable a
dversarial examples that are robust to the model variation, thus improving the s
uccess rate for attacking the black-box model. We conduct extensive experiments
on the ImageNet dataset with different classification models, and our proposed a
pproach outperforms existing state-of-the-art methods by a significant margin.
****
Efficient Loss Function by Minimizing the Detrimental Effect of Floating-Point E
rrors on Gradient-Based Attacks
Yunrui Yu, Cheng-Zhong Xu; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2023, pp. 4056-4066
Attackers can deceive neural networks by adding human imperceptive perturbations
 to their input data; this reveals the vulnerability and weak robustness of curr
ent deep-learning networks. Many attack techniques have been proposed to evaluat
e the model's robustness. Gradient-based attacks suffer from severely overestima
ting the robustness. This paper identifies that the relative error in calculated
 gradients caused by floating-point errors, including floating-point underflow a
nd rounding errors, is a fundamental reason why gradient-based attacks fail to a
ccurately assess the model's robustness. Although it is hard to eliminate the re
lative error in the gradients, we can control its effect on the gradient-based a
ttacks. Correspondingly, we propose an efficient loss function by minimizing the
 detrimental impact of the floating-point errors on the attacks. Experimental re
sults show that it is more efficient and reliable than other loss functions when
 examined across a wide range of defence mechanisms.
****

BAD-NeRF: Bundle Adjusted Deblur Neural Radiance Fields
Peng Wang, Lingzhe Zhao, Ruijie Ma, Peidong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4170-4179
Neural Radiance Fields (NeRF) have received considerable attention recently, due to its impressive capability in photo-realistic 3D reconstruction and novel view synthesis, given a set of posed camera images. Earlier work usually assumes the input images are of good quality. However, image degradation (e.g. image motion blur in low-light conditions) can easily happen in real-world scenarios, which would further affect the rendering quality of NeRF. In this paper, we present a novel bundle adjusted deblur Neural Radiance Fields (BAD-NeRF), which can be robust to severe motion blurred images and inaccurate camera poses. Our approach models the physical image formation process of a motion blurred image, and jointly learns the parameters of NeRF and recovers the camera motion trajectories during exposure time. In experiments, we show that by directly modeling the real physical image formation process, BAD-NeRF achieves superior performance over prior works on both synthetic and real datasets. Code and data are available at https://github.com/WU-CVGL/BAD-NeRF.
****
Video Compression With Entropy-Constrained Neural Representations
Carlos Gomes, Roberto Azevedo, Christopher Schroers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18497-18506
Encoding videos as neural networks is a recently proposed approach that allows new forms of video processing. However, traditional techniques still outperform such neural video representation (NVR) methods for the task of video compression. This performance gap can be explained by the fact that current NVR methods: i) use architectures that do not efficiently obtain a compact representation of temporal and spatial information; and ii) minimize rate and distortion disjointly (first overfitting a network on a video and then using heuristic techniques such as post-training quantization or weight pruning to compress the model). We propose a novel convolutional architecture for video representation that better represents spatio-temporal information and a training strategy capable of jointly optimizing rate and distortion. All network and quantization parameters are jointly learned end-to-end, and the post-training operations used in previous works are unnecessary. We evaluate our method on the UVG dataset, achieving new state-of-the-art results for video compression with NVRs. Moreover, we deliver the first NVR-based video compression method that improves over the typically adopted HEVC benchmark (x265, disabled b-frames, "medium" preset), closing the gap to autoencoder-based video compression techniques.
****
Prompt, Generate, Then Cache: Cascade of Foundation Models Makes Strong Few-Shot Learners
Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15211-15222
Visual recognition in low-data regimes requires deep neural networks to learn generalized representations from limited training samples. Recently, CLIP-based methods have shown promising few-shot performance benefited from the contrastive language-image pre-training. We then question, if the more diverse pre-training knowledge can be cascaded to further assist few-shot representation learning. In this paper, we propose CaFo, a Cascade of Foundation models that incorporates diverse prior knowledge of various pre training paradigms for better few-shot learning. Our CaFo incorporates CLIP's language-contrastive knowledge, DINO's vision-contrastive knowledge, DALL-E's vision generative knowledge, and GPT-3's language-generative knowledge. Specifically, CaFo works by 'Prompt, Generate, then Cache'. Firstly, we leverage GPT-3 to produce textual inputs for prompting CLIP with rich downstream linguistic semantics. Then, we generate synthetic images via DALL-E to expand the few-shot training data without any manpower. At last, we introduce a learnable cache model to adaptively blend the predictions from CLIP and DINO. By such col laboration, CaFo can fully unleash the potential of different

pre-training methods and unify them to perform state-of-the-art for few-shot classification. Code is available at https://github.com/ZrrSkywalker/CaFo.
****
Deep Random Projector: Accelerated Deep Image Prior
Taihui Li, Hengkang Wang, Zhong Zhuang, Ju Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18176-18185
Deep image prior (DIP) has shown great promise in tackling a variety of image restoration (IR) and general visual inverse problems, needing no training data. However, the resulting optimization process is often very slow, inevitably hindering DIP's practical usage for time-sensitive scenarios. In this paper, we focus on IR, and propose two crucial modifications to DIP that help achieve substantial speedup: 1) optimizing the DIP seed while freezing randomly-initialized network weights, and 2) reducing the network depth. In addition, we reintroduce explicit priors, such as sparse gradient prior---encoded by total-variation regularization, to preserve the DIP peak performance. We evaluate the proposed method on three IR tasks, including image denoising, image super-resolution, and image inpainting, against the original DIP and variants, as well as the competing metaDIP that uses meta-learning to learn good initializers with extra data. Our method is a clear winner in obtaining competitive restoration quality in a minimal amount of time. Our code is available at https://github.com/sun-umn/Deep-Random-Projector.
****
SCPNet: Semantic Scene Completion on Point Cloud
Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17642-17651
Training deep models for semantic scene completion is challenging due to the sparse and incomplete input, a large quantity of objects of diverse scales as well as the inherent label noise for moving objects. To address the above-mentioned problems, we propose the following three solutions: 1) Redesigning the completion network. We design a novel completion network, which consists of several Multi-Path Blocks (MPBs) to aggregate multi-scale features and is free from the lossy downsampling operations. 2) Distilling rich knowledge from the multi-frame model. We design a novel knowledge distillation objective, dubbed Dense-to-Sparse Knowledge Distillation (DSKD). It transfers the dense, relation-based semantic knowledge from the multi-frame teacher to the single-frame student, significantly improving the representation learning of the single-frame model. 3) Completion label rectification. We propose a simple yet effective label rectification strategy, which uses off-the-shelf panoptic segmentation labels to remove the traces of dynamic objects in completion labels, greatly improving the performance of deep models especially for those moving objects. Extensive experiments are conducted in two public semantic scene completion benchmarks, i.e., SemanticKITTI and SemanticPOSS. Our SCPNet ranks 1st on SemanticKITTI semantic scene completion challenge and surpasses the competitive S3CNet by 7.2 mIoU. SCPNet also outperforms previous completion algorithms on the SemanticPOSS dataset. Besides, our method also achieves competitive results on SemanticKITTI semantic segmentation tasks, showing that knowledge learned in the scene completion is beneficial to the segmentation task.
****
Revisiting Prototypical Network for Cross Domain Few-Shot Learning
Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20061-20070
Prototypical Network is a popular few-shot solver that aims at establishing a feature metric generalizable to novel few-shot classification (FSC) tasks using deep neural networks. However, its performance drops dramatically when generalizing to the FSC tasks in new domains. In this study, we revisit this problem and argue that the devil lies in the simplicity bias pitfall in neural networks. In specific, the network tends to focus on some biased shortcut features (e.g., color, shape, etc.) that are exclusively sufficient to distinguish very few classes i

n the meta-training tasks within a pre-defined domain, but fail to generalize ac
ross domains as some desirable semantic features. To mitigate this problem, we p
ropose a Local-global Distillation Prototypical Network (LDP-net). Different fro
m the standard Prototypical Network, we establish a two-branch network to classi
fy the query image and its random local crops, respectively. Then, knowledge dis
tillation is conducted among these two branches to enforce their class affiliati
on consistency. The rationale behind is that since such global-local semantic re
lationship is expected to hold regardless of data domains, the local-global dist
illation is beneficial to exploit some cross-domain transferable semantic featur
es for feature metric establishment. Moreover, such local-global semantic consis
tency is further enforced among different images of the same class to reduce the
 intra-class semantic variation of the resultant feature. In addition, we propos
e to update the local branch as Exponential Moving Average (EMA) over training e
pisodes, which makes it possible to better distill cross-episode knowledge and f
urther enhance the generalization performance. Experiments on eight cross-domain
 FSC benchmarks empirically clarify our argument and show the state-of-the-art r
esults of LDP-net. Code is available in https://github.com/NWPUZhoufei/LDP-Net
****
QPGesture: Quantization-Based and Phase-Guided Motion Matching for Natural Speec
h-Driven Gesture Generation
Sicheng Yang, Zhiyong Wu, Minglei Li, Zhensong Zhang, Lei Hao, Weihong Bao, Haol
in Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2023, pp. 2321-2330
Speech-driven gesture generation is highly challenging due to the random jitters
 of human motion. In addition, there is an inherent asynchronous relationship be
tween human speech and gestures. To tackle these challenges, we introduce a nove
l quantization-based and phase-guided motion matching framework. Specifically, w
e first present a gesture VQ-VAE module to learn a codebook to summarize meaning
ful gesture units. With each code representing a unique gesture, random jitterin
g problems are alleviated effectively. We then use Levenshtein distance to align
 diverse gestures with different speech. Levenshtein distance based on audio qua
ntization as a similarity metric of corresponding speech of gestures helps match
 more appropriate gestures with speech, and solves the alignment problem of spee
ch and gestures well. Moreover, we introduce phase to guide the optimal gesture
matching based on the semantics of context or rhythm of audio. Phase guides when
 text-based or speech-based gestures should be performed to make the generated g
estures more natural. Extensive experiments show that our method outperforms rec
ent approaches on speech-driven gesture generation. Our code, database, pre-trai
ned models and demos are available at https://github.com/YoungSeng/QPGesture.
****
Multiscale Tensor Decomposition and Rendering Equation Encoding for View Synthes
is
Kang Han, Wei Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision a
nd Pattern Recognition (CVPR), 2023, pp. 4232-4241
Rendering novel views from captured multi-view images has made considerable prog
ress since the emergence of the neural radiance field. This paper aims to furthe
r advance the quality of view rendering by proposing a novel approach dubbed the
 neural radiance feature field (NRFF). We first propose a multiscale tensor deco
mposition scheme to organize learnable features so as to represent scenes from c
oarse to fine scales. We demonstrate many benefits of the proposed multiscale re
presentation, including more accurate scene shape and appearance reconstruction,
 and faster convergence compared with the single-scale representation. Instead o
f encoding view directions to model view-dependent effects, we further propose t
o encode the rendering equation in the feature space by employing the anisotropi
c spherical Gaussian mixture predicted from the proposed multiscale representati
on. The proposed NRFF improves state-of-the-art rendering results by over 1 dB i
n PSNR on both the NeRF and NSVF synthetic datasets. A significant improvement h
as also been observed on the real-world Tanks & Temples dataset. Code can be fou
nd at https://github.com/imkanghan/nrff.
****

NS3D: Neuro-Symbolic Grounding of 3D Objects and Relations
Joy Hsu, Jiayuan Mao, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2614-2623
Grounding object properties and relations in 3D scenes is a prerequisite for a wide range of artificial intelligence tasks, such as visually grounded dialogues and embodied manipulation. However, the variability of the 3D domain induces two fundamental challenges: 1) the expense of labeling and 2) the complexity of 3D grounded language. Hence, essential desiderata for models are to be data-efficient, generalize to different data distributions and tasks with unseen semantic forms, as well as ground complex language semantics (e.g., view-point anchoring and multi-object reference). To address these challenges, we propose NS3D, a neuro-symbolic framework for 3D grounding. NS3D translates language into programs with hierarchical structures by leveraging large language-to-code models. Different functional modules in the programs are implemented as neural networks. Notably, NS3D extends prior neuro-symbolic visual reasoning methods by introducing functional modules that effectively reason about high-arity relations (i.e., relations among more than two objects), key in disambiguating objects in complex 3D scenes. Modular and compositional architecture enables NS3D to achieve state-of-the-art results on the ReferIt3D view-dependence task, a 3D referring expression comprehension benchmark. Importantly, NS3D shows significantly improved performance on settings of data-efficiency and generalization, and demonstrate zero-shot transfer to an unseen 3D question-answering task.
****
Learning Accurate 3D Shape Based on Stereo Polarimetric Imaging
Tianyu Huang, Haoang Li, Kejing He, Congying Sui, Bin Li, Yun-Hui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17287-17296
Shape from Polarization (SfP) aims to recover surface normal using the polarization cues of light. The accuracy of existing SfP methods is affected by two main problems. First, the ambiguity of polarization cues partially results in false normal estimation. Second, the widely-used assumption about orthographic projection is too ideal. To solve these problems, we propose the first approach that combines deep learning and stereo polarization information to recover not only normal but also disparity. Specifically, for the ambiguity problem, we design a Shape Consistency-based Mask Prediction (SCMP) module. It exploits the inherent consistency between normal and disparity to identify the areas with false normal estimation. We replace the unreliable features enclosed by these areas with new features extracted by global attention mechanism. As to the orthographic projection problem, we propose a novel Viewing Direction-aided Positional Encoding (VDPE) strategy. This strategy is based on the unique pixel-viewing direction encoding, and thus enables our neural network to handle the non-orthographic projection. In addition, we establish a real-world stereo SfP dataset that contains various object categories and illumination conditions. Experiments showed that compared with existing SfP methods, our approach is more accurate. Moreover, our approach shows higher robustness to light variation.
****
VideoMAE V2: Scaling Video Masked Autoencoders With Dual Masking
Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14549-14560
Scale is the primary factor for building a powerful foundation model that could well generalize to a variety of downstream tasks. However, it is still challenging to train video foundation models with billions of parameters. This paper shows that video masked autoencoder (VideoMAE) is a scalable and general self-supervised pre-trainer for building video foundation models. We scale the VideoMAE in both model and data with a core design. Specifically, we present a dual masking strategy for efficient pre-training, with an encoder operating on a subset of video tokens and a decoder processing another subset of video tokens. Although VideoMAE is very efficient due to high masking ratio in encoder, masking decoder can still further reduce the overall computational cost. This enables the efficien

t pre-training of billion-level models in video. We also introduce a progressive training paradigm that involves initial pre-training on the diverse multi-sourced unlabeled dataset, followed by fine-tuning on a mixed labeled dataset. Finally, we successfully train a video ViT model with a billion parameters, which achieves a new state-of-the-art performance on the datasets of Kinetics (90.0% on K400 and 89.9% on K600) and Something-Something (68.7% on V1 and 77.0% on V2). In addition, we extensively verify the pre-trained video ViT models on a variety of downstream tasks, demonstrating its effectiveness as a general video representation learner.
****
GANmouflage: 3D Object Nondetection With Texture Fields
Rui Guo, Jasmine Collins, Oscar de Lima, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4702-4712
We propose a method that learns to camouflage 3D objects within scenes. Given an object's shape and a distribution of viewpoints from which it will be seen, we estimate a texture that will make it difficult to detect. Successfully solving this task requires a model that can accurately reproduce textures from the scene, while simultaneously dealing with the highly conflicting constraints imposed by each viewpoint. We address these challenges with a model based on texture fields and adversarial learning. Our model learns to camouflage a variety of object shapes from randomly sampled locations and viewpoints within the input scene, and is the first to address the problem of hiding complex object shapes. Using a human visual search study, we find that our estimated textures conceal objects significantly better than previous methods.
****
Perception and Semantic Aware Regularization for Sequential Confidence Calibration
Zhenghua Peng, Yu Luo, Tianshui Chen, Keke Xu, Shuangping Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10658-10668
Deep sequence recognition (DSR) models receive increasing attention due to their superior application to various applications. Most DSR models use merely the target sequences as supervision without considering other related sequences, leading to over-confidence in their predictions. The DSR models trained with label smoothing regularize labels by equally and independently smoothing each token, reallocating a small value to other tokens for mitigating overconfidence. However, they do not consider tokens/sequences correlations that may provide more effective information to regularize training and thus lead to sub-optimal performance. In this work, we find tokens/sequences with high perception and semantic correlations with the target ones contain more correlated and effective information and thus facilitate more effective regularization. To this end, we propose a Perception and Semantic aware Sequence Regularization framework, which explore perceptively and semantically correlated tokens/sequences as regularization. Specifically, we introduce a semantic context-free recognition and a language model to acquire similar sequences with high perceptive similarities and semantic correlation, respectively. Moreover, over-confidence degree varies across samples according to their difficulties. Thus, we further design an adaptive calibration intensity module to compute a difficulty score for each samples to obtain finer-grained regularization. Extensive experiments on canonical sequence recognition tasks, including scene text and speech recognition, demonstrate that our method sets novel state-of-the-art results. Code is available at https://github.com/husterpzh/PSSR.
****
Revisiting Residual Networks for Adversarial Robustness
Shihua Huang, Zhichao Lu, Kalyanmoy Deb, Vishnu Naresh Boddeti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8202-8211
Efforts to improve the adversarial robustness of convolutional neural networks have primarily focused on developing more effective adversarial training methods.

In contrast, little attention was devoted to analyzing the role of architectural elements (e.g., topology, depth, and width) on adversarial robustness. This paper seeks to bridge this gap and present a holistic study on the impact of architectural design on adversarial robustness. We focus on residual networks and consider architecture design at the block level as well as at the network scaling level. In both cases, we first derive insights through systematic experiments. Then we design a robust residual block, dubbed RobustResBlock, and a compound scaling rule, dubbed RobustScaling, to distribute depth and width at the desired FLOP count. Finally, we combine RobustResBlock and RobustScaling and present a portfolio of adversarially robust residual networks, RobustResNets, spanning a broad spectrum of model capacities. Experimental validation across multiple datasets and adversarial attacks demonstrate that RobustResNets consistently outperform both the standard WRNs and other existing robust architectures, achieving state-of-the-art AutoAttack robust accuracy 63.7% with 500K external data while being 2x more compact in terms of parameters. The code is available at https://github.com/zhichao-lu/robust-residual-network.
****
Vision Transformer With Super Token Sampling
Huaibo Huang, Xiaoqiang Zhou, Jie Cao, Ran He, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22690-22699
Vision transformer has achieved impressive performance for many vision tasks. However, it may suffer from high redundancy in capturing local features for shallow layers. Local self-attention or early-stage convolutions are thus utilized, which sacrifice the capacity to capture long-range dependency. A challenge then arises: can we access efficient and effective global context modeling at the early stages of a neural network? To address this issue, we draw inspiration from the design of superpixels, which reduces the number of image primitives in subsequent processing, and introduce super tokens into vision transformer. Super tokens attempt to provide a semantically meaningful tessellation of visual content, thus reducing the token number in self-attention as well as preserving global modeling. Specifically, we propose a simple yet strong super token attention (STA) mechanism with three steps: the first samples super tokens from visual tokens via sparse association learning, the second performs self-attention on super tokens, and the last maps them back to the original token space. STA decomposes vanilla global attention into multiplications of a sparse association map and a low-dimensional attention, leading to high efficiency in capturing global dependencies. Based on STA, we develop a hierarchical vision transformer. Extensive experiments demonstrate its strong performance on various vision tasks. In particular, it achieves 86.4% top-1 accuracy on ImageNet-1K without any extra training data or label, 53.9 box AP and 46.8 mask AP on the COCO detection task, and 51.9 mIOU on the ADE20K semantic segmentation task.
****
RA-CLIP: Retrieval Augmented Contrastive Language-Image Pre-Training
Chen-Wei Xie, Siyang Sun, Xiong Xiong, Yun Zheng, Deli Zhao, Jingren Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19265-19274
Contrastive Language-Image Pre-training (CLIP) is attracting increasing attention for its impressive zero-shot recognition performance on different down-stream tasks. However, training CLIP is data-hungry and requires lots of image-text pairs to memorize various semantic concepts. In this paper, we propose a novel and efficient framework: Retrieval Augmented Contrastive Language-Image Pre-training (RA-CLIP) to augment embeddings by online retrieval. Specifically, we sample part of image-text data as a hold-out reference set. Given an input image, relevant image-text pairs are retrieved from the reference set to enrich the representation of input image. This process can be considered as an open-book exam: with the reference set as a cheat sheet, the proposed method doesn't need to memorize all visual concepts in the training data. It explores how to recognize visual concepts by exploiting correspondence between images and texts in the cheat sheet. The proposed RA-CLIP implements this idea and comprehensive experiments are con

ducted to show how RA-CLIP works. Performances on 10 image classification datasets and 2 object detection datasets show that RA-CLIP outperforms vanilla CLIP baseline by a large margin on zero-shot image classification task (+12.7%), linear probe image classification task (+6.9%) and zero-shot ROI classification task (+2.8%).
****
PosterLayout: A New Benchmark and Approach for Content-Aware Visual-Textual Presentation Layout
Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, Qing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6018-6026
Content-aware visual-textual presentation layout aims at arranging spatial space on the given canvas for pre-defined elements, including text, logo, and underlay, which is a key to automatic template-free creative graphic design. In practical applications, e.g., poster designs, the canvas is originally non-empty, and both inter-element relationships as well as inter-layer relationships should be concerned when generating a proper layout. A few recent works deal with them simultaneously, but they still suffer from poor graphic performance, such as a lack of layout variety or spatial non-alignment. Since content-aware visual-textual presentation layout is a novel task, we first construct a new dataset named PKU PosterLayout, which consists of 9,974 poster-layout pairs and 905 images, i.e., non-empty canvases. It is more challenging and useful for greater layout variety, domain diversity, and content diversity. Then, we propose design sequence formation (DSF) that reorganizes elements in layouts to imitate the design processes of human designers, and a novel CNN-LSTM-based conditional generative adversarial network (GAN) is presented to generate proper layouts. Specifically, the discriminator is design-sequence-aware and will supervise the "design" process of the generator. Experimental results verify the usefulness of the new benchmark and the effectiveness of the proposed approach, which achieves the best performance by generating suitable layouts for diverse canvases. The dataset and the source code are available at https://github.com/PKU-ICST-MIPL/PosterLayout-CVPR2023.
****
A Practical Upper Bound for the Worst-Case Attribution Deviations
Fan Wang, Adams Wai-Kin Kong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24616-24625
Model attribution is a critical component of deep neural networks (DNNs) for its interpretability to complex models. Recent studies bring up attention to the security of attribution methods as they are vulnerable to attribution attacks that generate similar images with dramatically different attributions. Existing works have been investigating empirically improving the robustness of DNNs against those attacks; however, none of them explicitly quantifies the actual deviations of attributions. In this work, for the first time, a constrained optimization problem is formulated to derive an upper bound that measures the largest dissimilarity of attributions after the samples are perturbed by any noises within a certain region while the classification results remain the same. Based on the formulation, different practical approaches are introduced to bound the attributions above using Euclidean distance and cosine similarity under both L2 and Linf-norm perturbations constraints. The bounds developed by our theoretical study are validated on various datasets and two different types of attacks (PGD attack and IFIA attribution attack). Over 10 million attacks in the experiments indicate that the proposed upper bounds effectively quantify the robustness of models based on the worst-case attribution dissimilarities.
****
A General Regret Bound of Preconditioned Gradient Method for DNN Training
Hongwei Yong, Ying Sun, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7866-7875
While adaptive learning rate methods, such as Adam, have achieved remarkable improvement in optimizing Deep Neural Networks (DNNs), they consider only the diagonal elements of the full preconditioned matrix. Though the full-matrix preconditioned gradient methods theoretically have a lower regret bound, they are impract

ical for use to train DNNs because of the high complexity. In this paper, we present a general regret bound with a constrained full-matrix preconditioned gradient and show that the updating formula of the preconditioner can be derived by solving a cone-constrained optimization problem. With the block-diagonal and Kronecker-factorized constraints, a specific guide function can be obtained. By minimizing the upper bound of the guide function, we develop a new DNN optimizer, termed AdaBK. A series of techniques, including statistics updating, dampening, efficient matrix inverse root computation, and gradient amplitude preservation, are developed to make AdaBK effective and efficient to implement. The proposed AdaBK can be readily embedded into many existing DNN optimizers, e.g., SGDM and AdamW, and the corresponding SGDM_BK and AdamW_BK algorithms demonstrate significant improvements over existing DNN optimizers on benchmark vision tasks, including image classification, object detection and segmentation. The source code will be made publicly available.
****
Teacher-Generated Spatial-Attention Labels Boost Robustness and Accuracy of Contrastive Models
Yushi Yao, Chang Ye, Junfeng He, Gamaleldin F. Elsayed; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23282-23291
Human spatial attention conveys information about theregions of visual scenes that are important for perform-ing visual tasks. Prior work has shown that the informa-tion about human attention can be leveraged to benefit var-ious supervised vision tasks. Might providing this weakform of supervision be useful for self-supervised represen-tation learning? Addressing this question requires collect-ing large datasets with human attention labels. Yet, col-lecting such large scale data is very expensive. To addressthis challenge, we construct an auxiliary teacher model topredict human attention, trained on a relatively small la-beled dataset. This teacher model allows us to generate im-age (pseudo) attention labels for ImageNet. We then traina model with a primary contrastive objective; to this stan-dard configuration, we add a simple output head trained topredict the attentional map for each image, guided by thepseudo labels from teacher model. We measure the qual-ity of learned representations by evaluating classificationperformance from the frozen learned embeddings as wellas performance on image retrieval tasks. We find that thespatial-attention maps predicted from the contrastive modeltrained with teacher guidance aligns better with human at-tention compared to vanilla contrastive models. Moreover,we find that our approach improves classification accuracyand robustness of the contrastive models on ImageNet andImageNet-C. Further, we find that model representationsbecome more useful for image retrieval task as measuredby precision-recall performance on ImageNet, ImageNet-C,CIFAR10, and CIFAR10-C datasets.
****
Exploring and Exploiting Uncertainty for Incomplete Multi-View Classification
Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19873-19882
Classifying incomplete multi-view data is inevitable since arbitrary view missing widely exists in real-world applications. Although great progress has been achieved, existing incomplete multi-view methods are still difficult to obtain a trustworthy prediction due to the relatively high uncertainty nature of missing views. First, the missing view is of high uncertainty, and thus it is not reasonable to provide a single deterministic imputation. Second, the quality of the imputed data itself is of high uncertainty. To explore and exploit the uncertainty, we propose an Uncertainty-induced Incomplete Multi-View Data Classification (UIMC) model to classify the incomplete multi-view data under a stable and reliable framework. We construct a distribution and sample multiple times to characterize the uncertainty of missing views, and adaptively utilize them according to the sampling quality. Accordingly, the proposed method realizes more perceivable imputation and controllable fusion. Specifically, we model each missing data with a distribution conditioning on the available views and thus introducing uncertain

ty. Then an evidence-based fusion strategy is employed to guarantee the trustwor thy integration of the imputed views. Extensive experiments are conducted on mul tiple benchmark data sets and our method establishes a state-of-the-art performa nce in terms of both performance and trustworthiness.
****
Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Capt ioning
Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, Cordelia Schmid; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10714-10726
In this work, we introduce Vid2Seq, a multi-modal single-stage dense event capti oning model pretrained on narrated videos which are readily-available at scale. The Vid2Seq architecture augments a language model with special time tokens, all owing it to seamlessly predict event boundaries and textual descriptions in the same output sequence. Such a unified model requires large-scale training data, w hich is not available in current annotated datasets. We show that it is possible to leverage unlabeled narrated videos for dense video captioning, by reformulat ing sentence boundaries of transcribed speech as pseudo event boundaries, and us ing the transcribed speech sentences as pseudo event captions. The resulting Vid 2Seq model pretrained on the YT-Temporal-1B dataset improves the state of the ar t on a variety of dense video captioning benchmarks including YouCook2, ViTT and ActivityNet Captions. Vid2Seq also generalizes well to the tasks of video parag raph captioning and video clip captioning, and to few-shot settings. Our code is publicly available at https://antoyang.github.io/vid2seq.html.
****
Optimal Proposal Learning for Deployable End-to-End Pedestrian Detection
Xiaolin Song, Binghui Chen, Pengyu Li, Jun-Yan He, Biao Wang, Yifeng Geng, Xuans ong Xie, Honggang Zhang; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2023, pp. 3250-3260
End-to-end pedestrian detection focuses on training a pedestrian detection model via discarding the Non-Maximum Suppression (NMS) post-processing. Though a few methods have been explored, most of them still suffer from longer training time and more complex deployment, which cannot be deployed in the actual industrial a pplications. In this paper, we intend to bridge this gap and propose an Optimal Proposal Learning (OPL) framework for deployable end-to-end pedestrian detection . Specifically, we achieve this goal by using CNN-based light detector and intro ducing two novel modules, including a Coarse-to-Fine (C2F) learning strategy for proposing precise positive proposals for the Ground-Truth (GT) instances by red ucing the ambiguity of sample assignment/output in training/testing respectively , and a Completed Proposal Network (CPN) for producing extra information compens ation to further recall the hard pedestrian samples. Extensive experiments are c onducted on CrowdHuman, TJU-Ped and Caltech, and the results show that our propo sed OPL method significantly outperforms the competing methods.
****
Discovering the Real Association: Multimodal Causal Reasoning in Video Question Answering
Chuanqi Zang, Hanqing Wang, Mingtao Pei, Wei Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19027-19 036
Video Question Answering (VideoQA) is challenging as it requires capturing accur ate correlations between modalities from redundant information. Recent methods f ocus on the explicit challenges of the task, e.g. multimodal feature extraction, video-text alignment and fusion. Their frameworks reason the answer relying on statistical evidence causes, which ignores potential bias in the multimodal data . In our work, we investigate relational structure from a causal representation perspective on multimodal data and propose a novel inference framework. For visu al data, question-irrelevant objects may establish simple matching associations with the answer. For textual data, the model prefers the local phrase semantics which may deviate from the global semantics in long sentences. Therefore, to enh ance the generalization of the model, we discover the real association by explic

itly capturing visual features that are causally related to the question semantics and weakening the impact of local language semantics on question answering. The experimental results on two large causal VideoQA datasets verify that our proposed framework 1) improves the accuracy of the existing VideoQA backbone, 2) demonstrates robustness on complex scenes and questions.
****

Temporal Interpolation Is All You Need for Dynamic Neural Radiance Fields
Sungheon Park, Minjung Son, Seokhwan Jang, Young Chun Ahn, Ji-Yeon Kim, Nahyup Kang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4212-4221
Temporal interpolation often plays a crucial role to learn meaningful representations in dynamic scenes. In this paper, we propose a novel method to train spatiotemporal neural radiance fields of dynamic scenes based on temporal interpolation of feature vectors. Two feature interpolation methods are suggested depending on underlying representations, neural networks or grids. In the neural representation, we extract features from space-time inputs via multiple neural network modules and interpolate them based on time frames. The proposed multi-level feature interpolation network effectively captures features of both short-term and long-term time ranges. In the grid representation, space-time features are learned via four-dimensional hash grids, which remarkably reduces training time. The grid representation shows more than 100 times faster training speed than the previous neural-net-based methods while maintaining the rendering quality. Concatenating static and dynamic features and adding a simple smoothness term further improve the performance of our proposed models. Despite the simplicity of the model architectures, our method achieved state-of-the-art performance both in rendering quality for the neural representation and in training speed for the grid representation.
****

Graph Transformer GANs for Graph-Constrained House Generation
Hao Tang, Zhenyu Zhang, Humphrey Shi, Bo Li, Ling Shao, Nicu Sebe, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2173-2182
We present a novel graph Transformer generative adversarial network (GTGAN) to learn effective graph node relations in an end-to-end fashion for the challenging graph-constrained house generation task. The proposed graph-Transformer-based generator includes a novel graph Transformer encoder that combines graph convolutions and self-attentions in a Transformer to model both local and global interactions across connected and non-connected graph nodes. Specifically, the proposed connected node attention (CNA) and non-connected node attention (NNA) aim to capture the global relations across connected nodes and non-connected nodes in the input graph, respectively. The proposed graph modeling block (GMB) aims to exploit local vertex interactions based on a house layout topology. Moreover, we propose a new node classification-based discriminator to preserve the high-level semantic and discriminative node features for different house components. Finally, we propose a novel graph-based cycle-consistency loss that aims at maintaining the relative spatial relationships between ground truth and predicted graphs. Experiments on two challenging graph-constrained house generation tasks (i.e., house layout and roof generation) with two public datasets demonstrate the effectiveness of GTGAN in terms of objective quantitative scores and subjective visual realism. New state-of-the-art results are established by large margins on both tasks.
****

On the Benefits of 3D Pose and Tracking for Human Action Recognition
Jathushan Rajasegaran, Georgios Pavlakos, Angjoo Kanazawa, Christoph Feichtenhofer, Jitendra Malik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 640-649
In this work we study the benefits of using tracking and 3D poses for action recognition. To achieve this, we take the Lagrangian view on analysing actions over a trajectory of human motion rather than at a fixed point in space. Taking this stand allows us to use the tracklets of people to predict their actions. In thi

s spirit, first we show the benefits of using 3D pose to infer actions, and study person-person interactions. Subsequently, we propose a Lagrangian Action Recognition model by fusing 3D pose and contextualized appearance over tracklets. To this end, our method achieves state-of-the-art performance on the AVA v2.2 dataset on both pose only settings and on standard benchmark settings. When reasoning about the action using only pose cues, our pose model achieves +10.0 mAP gain over the corresponding state-of-the-art while our fused model has a gain of +2.8 mAP over the best state-of-the-art model. Code and results are available at: https://brjathu.github.io/LART
****
How to Backdoor Diffusion Models?
Sheng-Yen Chou, Pin-Yu Chen, Tsung-Yi Ho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4015-4024
Diffusion models are state-of-the-art deep learning empowered generative models that are trained based on the principle of learning forward and reverse diffusion processes via progressive noise-addition and denoising. To gain a better understanding of the limitations and potential risks, this paper presents the first study on the robustness of diffusion models against backdoor attacks. Specifically, we propose BadDiffusion, a novel attack framework that engineers compromised diffusion processes during model training for backdoor implantation. At the inference stage, the backdoored diffusion model will behave just like an untampered generator for regular data inputs, while falsely generating some targeted outcome designed by the bad actor upon receiving the implanted trigger signal. Such a critical risk can be dreadful for downstream tasks and applications built upon the problematic model. Our extensive experiments on various backdoor attack settings show that BadDiffusion can consistently lead to compromised diffusion models with high utility and target specificity. Even worse, BadDiffusion can be made cost-effective by simply finetuning a clean pre-trained diffusion model to implant backdoors. We also explore some possible countermeasures for risk mitigation. Our results call attention to potential risks and possible misuse of diffusion models.
****
ERNIE-ViLG 2.0: Improving Text-to-Image Diffusion Model With Knowledge-Enhanced Mixture-of-Denoising-Experts
Zhida Feng, Zhenyu Zhang, Xintong Yu, Yewei Fang, Lanxin Li, Xuyi Chen, Yuxiang Lu, Jiaxiang Liu, Weichong Yin, Shikun Feng, Yu Sun, Li Chen, Hao Tian, Hua Wu, Haifeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10135-10145
Recent progress in diffusion models has revolutionized the popular technology of text-to-image generation. While existing approaches could produce photorealistic high-resolution images with text conditions, there are still several open problems to be solved, which limits the further improvement of image fidelity and text relevancy. In this paper, we propose ERNIE-ViLG 2.0, a large-scale Chinese text-to-image diffusion model, to progressively upgrade the quality of generated images by: (1) incorporating fine-grained textual and visual knowledge of key elements in the scene, and (2) utilizing different denoising experts at different denoising stages. With the proposed mechanisms, ERNIE-ViLG 2.0 not only achieves a new state-of-the-art on MS-COCO with zero-shot FID-30k score of 6.75, but also significantly outperforms recent models in terms of image fidelity and image-text alignment, with side-by-side human evaluation on the bilingual prompt set ViLG-300.
****
PACO: Parts and Attributes of Common Objects
Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, Amir Mousavi, Yiwen Song, Abhimanyu Dubey, Dhruv Mahajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7141-7151
Object models are gradually progressing from predicting just category labels to providing detailed descriptions of object instances. This motivates the need for large datasets which go beyond traditional object masks and provide richer anno

tations such as part masks and attributes. Hence, we introduce PACO: Parts and A ttributes of Common Objects. It spans 75 object categories, 456 object-part cate gories and 55 attributes across image (LVIS) and video (Ego4D) datasets. We prov ide 641K part masks annotated across 260K object boxes, with roughly half of the m exhaustively annotated with attributes as well. We design evaluation metrics a nd provide benchmark results for three tasks on the dataset: part mask segmentat ion, object and part attribute prediction and zero-shot instance detection. Data set, models, and code are open-sourced at https://github.com/facebookresearch/pa co.
****
Learning Transformations To Reduce the Geometric Shift in Object Detection
Vidit Vidit, Martin Engilberge, Mathieu Salzmann; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17441-1745 0
The performance of modern object detectors drops when the test distribution diff ers from the training one. Most of the methods that address this focus on object appearance changes caused by, e.g., different illumination conditions, or gaps between synthetic and real images. Here, by contrast, we tackle geometric shifts emerging from variations in the image capture process, or due to the constraint s of the environment causing differences in the apparent geometry of the content itself. We introduce a self-training approach that learns a set of geometric tr ansformations to minimize these shifts without leveraging any labeled data in th e new domain, nor any information about the cameras. We evaluate our method on t wo different shifts, i.e., a camera's field of view (FoV) change and a viewpoint change. Our results evidence that learning geometric transformations helps dete ctors to perform better in the target domains.
****
OReX: Object Reconstruction From Planar Cross-Sections Using Neural Fields
Haim Sawdayee, Amir Vaxman, Amit H. Bermano; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20854-20862
Reconstructing 3D shapes from planar cross-sections is a challenge inspired by d ownstream applications like medical imaging and geographic informatics. The inpu t is an in/out indicator function fully defined on a sparse collection of planes in space, and the output is an interpolation of the indicator function to the e ntire volume. Previous works addressing this sparse and ill-posed problem either produce low quality results, or rely on additional priors such as target topolo gy, appearance information, or input normal directions. In this paper, we presen t OReX, a method for 3D shape reconstruction from slices alone, featuring a Neur al Field as the interpolation prior. A modest neural network is trained on the i nput planes to return an inside/outside estimate for a given 3D coordinate, yiel ding a powerful prior that induces smoothness and self-similarities. The main ch allenge for this approach is high-frequency details, as the neural prior is over ly smoothing. To alleviate this, we offer an iterative estimation architecture a nd a hierarchical input sampling scheme that encourage coarse-to-fine training, allowing the training process to focus on high frequencies at later stages. In a ddition, we identify and analyze a ripple-like effect stemming from the mesh ext raction step. We mitigate it by regularizing the spatial gradients of the indica tor function around input in/out boundaries during network training, tackling th e problem at the root. Through extensive qualitative and quantitative experiment ation, we demonstrate our method is robust, accurate, and scales well with the s ize of the input. We report state-of-the-art results compared to previous approa ches and recent potential solutions, and demonstrate the benefit of our individu al contributions through analysis and ablation studies.
****
SPIn-NeRF: Multiview Segmentation and Perceptual Inpainting With Neural Radiance Fields
Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G. Derpanis, Jonathan Kelly, Marcus A. Brubaker, Igor Gilitschenski, Alex Levinshtein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20669-20679

Neural Radiance Fields (NeRFs) have emerged as a popular approach for novel view synthesis. While NeRFs are quickly being adapted for a wider set of applications, intuitively editing NeRF scenes is still an open challenge. One important editing task is the removal of unwanted objects from a 3D scene, such that the replaced region is visually plausible and consistent with its context. We refer to this task as 3D inpainting. In 3D, solutions must be both consistent across multiple views and geometrically valid. In this paper, we propose a novel 3D inpainting method that addresses these challenges. Given a small set of posed images and sparse annotations in a single input image, our framework first rapidly obtains a 3D segmentation mask for a target object. Using the mask, a perceptual optimization-based approach is then introduced that leverages learned 2D image inpainters, distilling their information into 3D space, while ensuring view consistency. We also address the lack of a diverse benchmark for evaluating 3D scene inpainting methods by introducing a dataset comprised of challenging real-world scenes. In particular, our dataset contains views of the same scene with and without a target object, enabling more principled benchmarking of the 3D inpainting task. We first demonstrate the superiority of our approach on multiview segmentation, comparing to NeRF-based methods and 2D segmentation approaches. We then evaluate on the task of 3D inpainting, establishing state-of-the-art performance against other NeRF manipulation algorithms, as well as a strong 2D image inpainter baseline.

****

Revisiting the Stack-Based Inverse Tone Mapping

Ning Zhang, Yuyao Ye, Yang Zhao, Ronggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9162-9171

Current stack-based inverse tone mapping (ITM) methods can recover high dynamic range (HDR) radiance by predicting a set of multi-exposure images from a single low dynamic range image. However, there are still some limitations. On the one hand, these methods estimate a fixed number of images (e.g., three exposure-up and three exposure-down), which may introduce unnecessary computational cost or reconstruct incorrect results. On the other hand, they neglect the connections between the up-exposure and down-exposure models and thus fail to fully excavate effective features. In this paper, we revisit the stack-based ITM approaches and propose a novel method to reconstruct HDR radiance from a single image, which only needs to estimate two exposure images. At first, we design the exposure adaptive block that can adaptively adjust the exposure based on the luminance distribution of the input image. Secondly, we devise the cross-model attention block to connect the exposure adjustment models. Thirdly, we propose an end-to-end ITM pipeline by incorporating the multi-exposure fusion model. Furthermore, we propose and open a multi-exposure dataset that indicates the optimal exposure-up/down levels. Experimental results show that the proposed method outperforms some state-of-the-art methods.

****

Revisiting Rotation Averaging: Uncertainties and Robust Losses

Ganlin Zhang, Viktor Larsson, Daniel Barath; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17215-17224

In this paper, we revisit the rotation averaging problem applied in global Structure-from-Motion pipelines. We argue that the main problem of current methods is the minimized cost function that is only weakly connected with the input data via the estimated epipolar geometries. We propose to better model the underlying noise distributions by directly propagating the uncertainty from the point correspondences into the rotation averaging. Such uncertainties are obtained for free by considering the Jacobians of two-view refinements. Moreover, we explore integrating a variant of the MAGSAC loss into the rotation averaging problem, instead of using classical robust losses employed in current frameworks. The proposed method leads to results superior to baselines, in terms of accuracy, on large-scale public benchmarks. The code is public. https://github.com/zhangganlin/GlobalSfMpy

****

Continuous Sign Language Recognition With Correlation Network

Lianyu Hu, Liqing Gao, Zekang Liu, Wei Feng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2529-2539
Human body trajectories are a salient cue to identify actions in video. Such body trajectories are mainly conveyed by hands and face across consecutive frames in sign language. However, current methods in continuous sign language recognition(CSLR) usually process frames independently to capture frame-wise features, thus failing to capture cross-frame trajectories to effectively identify a sign. To handle this limitation, we propose correlation network (CorrNet) to explicitly leverage body trajectories across frames to identify signs. In specific, an identification module is first presented to emphasize informative regions in each frame that are beneficial in expressing a sign. A correlation module is then proposed to dynamically compute correlation maps between current frame and adjacent neighbors to capture cross-frame trajectories. As a result, the generated features are able to gain an overview of local temporal movements to identify a sign. Thanks to its special attention on body trajectories, CorrNet achieves new state-of-the-art accuracy on four large-scale datasets, PHOENIX14, PHOENIX14-T, CSL-Daily, and CSL. A comprehensive comparison between CorrNet and previous spatial-temporal reasoning methods verifies its effectiveness. Visualizations are given to demonstrate the effects of CorrNet on emphasizing human body trajectories across adjacent frames.
****
A Simple Framework for Text-Supervised Semantic Segmentation
Muyang Yi, Quan Cui, Hao Wu, Cheng Yang, Osamu Yoshie, Hongtao Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7071-7080
Text-supervised semantic segmentation is a novel research topic that allows semantic segments to emerge with image-text contrasting. However, pioneering methods could be subject to specifically designed network architectures. This paper shows that a vanilla contrastive language-image pre-training (CLIP) model is an effective text-supervised semantic segmentor by itself. First, we reveal that a vanilla CLIP is inferior to localization and segmentation due to its optimization being driven by densely aligning visual and language representations. Second, we propose the locality-driven alignment (LoDA) to address the problem, where CLIP optimization is driven by sparsely aligning local representations. Third, we propose a simple segmentation (SimSeg) framework. LoDA and SimSeg jointly ameliorate a vanilla CLIP to produce impressive semantic segmentation results. Our method outperforms previous state-of-the-art methods on PASCAL VOC 2012, PASCAL Context and COCO datasets by large margins. Code and models are available at github.com/muyangyi/SimSeg.
****
Exploiting Completeness and Uncertainty of Pseudo Labels for Weakly Supervised Video Anomaly Detection
Chen Zhang, Guorong Li, Yuankai Qi, Shuhui Wang, Laiyun Qing, Qingming Huang, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16271-16280
Weakly supervised video anomaly detection aims to identify abnormal events in videos using only video-level labels. Recently, two-stage self-training methods have achieved significant improvements by self-generating pseudo labels and self-refining anomaly scores with these labels. As the pseudo labels play a crucial role, we propose an enhancement framework by exploiting completeness and uncertainty properties for effective self-training. Specifically, we first design a multi-head classification module (each head serves as a classifier) with a diversity loss to maximize the distribution differences of predicted pseudo labels across heads. This encourages the generated pseudo labels to cover as many abnormal events as possible. We then devise an iterative uncertainty pseudo label refinement strategy, which improves not only the initial pseudo labels but also the updated ones obtained by the desired classifier in the second stage. Extensive experimental results demonstrate the proposed method performs favorably against state-of-the-art approaches on the UCF-Crime, TAD, and XD-Violence benchmark datasets.
****

PlenVDB: Memory Efficient VDB-Based Radiance Fields for Fast Training and Rendering
Han Yan, Celong Liu, Chao Ma, Xing Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 88-96
In this paper, we present a new representation for neural radiance fields that accelerates both the training and the inference processes with VDB, a hierarchical data structure for sparse volumes. VDB takes both the advantages of sparse and dense volumes for compact data representation and efficient data access, being a promising data structure for NeRF data interpolation and ray marching. Our method, Plenoptic VDB (PlenVDB), directly learns the VDB data structure from a set of posed images by means of a novel training strategy and then uses it for real-time rendering. Experimental results demonstrate the effectiveness and the efficiency of our method over previous arts: First, it converges faster in the training process. Second, it delivers a more compact data format for NeRF data presentation. Finally, it renders more efficiently on commodity graphics hardware. Our mobile PlenVDB demo achieves 30+ FPS, 1280x720 resolution on an iPhone12 mobile phone. Check plenvdb.github.io for details.
****
Patch-Based 3D Natural Scene Generation From a Single Example
Weiyu Li, Xuelin Chen, Jue Wang, Baoquan Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16762-16772
We target a 3D generative model for general natural scenes that are typically unique and intricate. Lacking the necessary volumes of training data, along with the difficulties of having ad hoc designs in presence of varying scene characteristics, renders existing setups intractable. Inspired by classical patch-based image models, we advocate for synthesizing 3D scenes at the patch level, given a single example. At the core of this work lies important algorithmic designs w.r.t the scene representation and generative patch nearest-neighbor module, that address unique challenges arising from lifting classical 2D patch-based framework to 3D generation. These design choices, on a collective level, contribute to a robust, effective, and efficient model that can generate high-quality general natural scenes with both realistic geometric structure and visual appearance, in large quantities and varieties, as demonstrated upon a variety of exemplar scenes. Data and code can be found at http://wyysf-98.github.io/Sin3DGen.
****
Full or Weak Annotations? An Adaptive Strategy for Budget-Constrained Annotation Campaigns
Javier Gamazo Tejero, Martin S. Zinkernagel, Sebastian Wolf, Raphael Sznitman, Pablo Márquez-Neila; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11381-11391
Annotating new datasets for machine learning tasks is tedious, time-consuming, and costly. For segmentation applications, the burden is particularly high as manual delineations of relevant image content are often extremely expensive or can only be done by experts with domain-specific knowledge. Thanks to developments in transfer learning and training with weak supervision, segmentation models can now also greatly benefit from annotations of different kinds. However, for any new domain application looking to use weak supervision, the dataset builder still needs to define a strategy to distribute full segmentation and other weak annotations. Doing so is challenging, however, as it is a priori unknown how to distribute an annotation budget for a given new dataset. To this end, we propose a novel approach to determine annotation strategies for segmentation datasets, whereby estimating what proportion of segmentation and classification annotations should be collected given a fixed budget. To do so, our method sequentially determines proportions of segmentation and classification annotations to collect for budget-fractions by modeling the expected improvement of the final segmentation model. We show in our experiments that our approach yields annotations that perform very close to the optimal for a number of different annotation budgets and datasets.
****
Leveraging Hidden Positives for Unsupervised Semantic Segmentation

Hyun Seok Seong, WonJun Moon, SuBeen Lee, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19540-19549

Dramatic demand for manpower to label pixel-level annotations triggered the advent of unsupervised semantic segmentation. Although the recent work employing the vision transformer (ViT) backbone shows exceptional performance, there is still a lack of consideration for task-specific training guidance and local semantic consistency. To tackle these issues, we leverage contrastive learning by excavating hidden positives to learn rich semantic relationships and ensure semantic consistency in local regions. Specifically, we first discover two types of global hidden positives, task-agnostic and task-specific ones for each anchor based on the feature similarities defined by a fixed pre-trained backbone and a segmentation head-in-training, respectively. A gradual increase in the contribution of the latter induces the model to capture task-specific semantic features. In addition, we introduce a gradient propagation strategy to learn semantic consistency between adjacent patches, under the inherent premise that nearby patches are highly likely to possess the same semantics. Specifically, we add the loss propagating to local hidden positives, semantically similar nearby patches, in proportion to the predefined similarity scores. With these training schemes, our proposed method achieves new state-of-the-art (SOTA) results in COCO-stuff, Cityscapes, and Potsdam-3 datasets. Our code is available at: https://github.com/hynnsk/HP.

****

Backdoor Defense via Deconfounded Representation Learning

Zaixi Zhang, Qi Liu, Zhicai Wang, Zepu Lu, Qingyong Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12228-12238

Deep neural networks (DNNs) are recently shown to be vulnerable to backdoor attacks, where attackers embed hidden backdoors in the DNN model by injecting a few poisoned examples into the training dataset. While extensive efforts have been made to detect and remove backdoors from backdoored DNNs, it is still not clear whether a backdoor-free clean model can be directly obtained from poisoned datasets. In this paper, we first construct a causal graph to model the generation process of poisoned data and find that the backdoor attack acts as the confounder, which brings spurious associations between the input images and target labels, making the model predictions less reliable. Inspired by the causal understanding, we propose the Causality-inspired Backdoor Defense (CBD), to learn deconfounded representations by employing the front-door adjustment. Specifically, a backdoored model is intentionally trained to capture the confounding effects. The other clean model dedicates to capturing the desired causal effects by minimizing the mutual information with the confounding representations from the backdoored model and employing a sample-wise re-weighting scheme. Extensive experiments on multiple benchmark datasets against 6 state-of-the-art attacks verify that our proposed defense method is effective in reducing backdoor threats while maintaining high accuracy in predicting benign samples. Further analysis shows that CBD can also resist potential adaptive attacks.

****

LG-BPN: Local and Global Blind-Patch Network for Self-Supervised Real-World Denoising

Zichun Wang, Ying Fu, Ji Liu, Yulun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18156-18165

Despite the significant results on synthetic noise under simplified assumptions, most self-supervised denoising methods fail under real noise due to the strong spatial noise correlation, including the advanced self-supervised blind-spot networks (BSNs). For recent methods targeting real-world denoising, they either suffer from ignoring this spatial correlation, or are limited by the destruction of fine textures for under-considering the correlation. In this paper, we present a novel method called LG-BPN for self-supervised real-world denoising, which takes the spatial correlation statistic into our network design for local detail restoration, and also brings the long-range dependencies modeling ability to previously CNN-based BSN methods. First, based on the correlation statistic, we propo

se a densely-sampled patch-masked convolution module. By taking more neighbor pixels with low noise correlation into account, we enable a denser local receptive field, preserving more useful information for enhanced fine structure recovery. Second, we propose a dilated Transformer block to allow distant context exploitation in BSN. This global perception addresses the intrinsic deficiency of BSN, whose receptive field is constrained by the blind spot requirement, which can not be fully resolved by the previous CNN-based BSNs. These two designs enable LG-BPN to fully exploit both the detailed structure and the global interaction in a blind manner. Extensive results on real-world datasets demonstrate the superior performance of our method. https://github.com/Wang-XIaoDingdd/LGBPN
****
Efficient View Synthesis and 3D-Based Multi-Frame Denoising With Multiplane Feature Representations
Thomas Tanay, Aleš Leonardis, Matteo Maggioni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20898-20907
While current multi-frame restoration methods combine information from multiple input images using 2D alignment techniques, recent advances in novel view synthesis are paving the way for a new paradigm relying on volumetric scene representations. In this work, we introduce the first 3D-based multi-frame denoising method that significantly outperforms its 2D-based counterparts with lower computational requirements. Our method extends the multiplane image (MPI) framework for novel view synthesis by introducing a learnable encoder-renderer pair manipulating multiplane representations in feature space. The encoder fuses information across views and operates in a depth-wise manner while the renderer fuses information across depths and operates in a view-wise manner. The two modules are trained end-to-end and learn to separate depths in an unsupervised way, giving rise to Multiplane Feature (MPF) representations. Experiments on the Spaces and Real Forward-Facing datasets as well as on raw burst data validate our approach for view synthesis, multi-frame denoising, and view synthesis under noisy conditions.
****
An Actor-Centric Causality Graph for Asynchronous Temporal Inference in Group Activity
Zhao Xie, Tian Gao, Kewei Wu, Jiao Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6652-6661
The causality relation modeling remains a challenging task for group activity recognition. The causality relations describe the influence of some actors (cause actors) on other actors (effect actors). Most existing graph models focus on learning the actor relation with synchronous temporal features, which is insufficient to deal with the causality relation with asynchronous temporal features. In this paper, we propose an Actor-Centric Causality Graph Model, which learns the asynchronous temporal causality relation with three modules, i.e., an asynchronous temporal causality relation detection module, a causality feature fusion module, and a causality relation graph inference module. First, given a centric actor and correlative actor, we analyze their influences to detect causality relation. We estimate the self influence of the centric actor with self regression. We estimate the correlative influence from the correlative actor to the centric actor with correlative regression, which uses asynchronous features at different timestamps. Second, we synchronize the two action features by estimating the temporal delay between the cause action and the effect action. The synchronized features are used to enhance the feature of the effect action with a channel-wise fusion. Third, we describe the nodes (actors) with causality features and learn the edges by fusing the causality relation with the appearance relation and distance relation. The causality relation graph inference provides crucial features of effect action, which are complementary to the base model using synchronous relation inference. The two relation inferences are aggregated to enhance group relation learning. Extensive experiments show that our method achieves state-of-the-art performance on the Volleyball dataset and Collective Activity dataset.
****
Color Backdoor: A Robust Poisoning Attack in Color Space
Wenbo Jiang, Hongwei Li, Guowen Xu, Tianwei Zhang; Proceedings of the IEEE/CVF C

onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8133-8142
Backdoor attacks against neural networks have been intensively investigated, where the adversary compromises the integrity of the victim model, causing it to make wrong predictions for inference samples containing a specific trigger. To make the trigger more imperceptible and human-unnoticeable, a variety of stealthy backdoor attacks have been proposed, some works employ imperceptible perturbations as the backdoor triggers, which restrict the pixel differences of the triggered image and clean image. Some works use special image styles (e.g., reflection, Instagram filter) as the backdoor triggers. However, these attacks sacrifice the robustness, and can be easily defeated by common preprocessing-based defenses. This paper presents a novel color backdoor attack, which can exhibit robustness and stealthiness at the same time. The key insight of our attack is to apply a uniform color space shift for all pixels as the trigger. This global feature is robust to image transformation operations and the triggered samples maintain natural-looking. To find the optimal trigger, we first define naturalness restrictions through the metrics of PSNR, SSIM and LPIPS. Then we employ the Particle Swarm Optimization (PSO) algorithm to search for the optimal trigger that can achieve high attack effectiveness and robustness while satisfying the restrictions. Extensive experiments demonstrate the superiority of PSO and the robustness of color backdoor against different mainstream backdoor defenses.
****
HairStep: Transfer Synthetic to Real Using Strand and Depth Maps for Single-View 3D Hair Modeling
Yujian Zheng, Zirong Jin, Moran Li, Haibin Huang, Chongyang Ma, Shuguang Cui, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12726-12735
In this work, we tackle the challenging problem of learning-based single-view 3D hair modeling. Due to the great difficulty of collecting paired real image and 3D hair data, using synthetic data to provide prior knowledge for real domain becomes a leading solution. This unfortunately introduces the challenge of domain gap. Due to the inherent difficulty of realistic hair rendering, existing methods typically use orientation maps instead of hair images as input to bridge the gap. We firmly think an intermediate representation is essential, but we argue that orientation map using the dominant filtering-based methods is sensitive to uncertain noise and far from a competent representation. Thus, we first raise this issue up and propose a novel intermediate representation, termed as HairStep, which consists of a strand map and a depth map. It is found that HairStep not only provides sufficient information for accurate 3D hair modeling, but also is feasible to be inferred from real images. Specifically, we collect a dataset of 1,250 portrait images with two types of annotations. A learning framework is further designed to transfer real images to the strand map and depth map. It is noted that, an extra bonus of our new dataset is the first quantitative metric for 3D hair modeling. Our experiments show that HairStep narrows the domain gap between synthetic and real and achieves state-of-the-art performance on single-view 3D hair reconstruction.
****
MoDAR: Using Motion Forecasting for 3D Object Detection in Point Cloud Sequences
Yingwei Li, Charles R. Qi, Yin Zhou, Chenxi Liu, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9329-9339
Occluded and long-range objects are ubiquitous and challenging for 3D object detection. Point cloud sequence data provide unique opportunities to improve such cases, as an occluded or distant object can be observed from different viewpoints or gets better visibility over time. However, the efficiency and effectiveness in encoding long-term sequence data can still be improved. In this work, we propose MoDAR, using motion forecasting outputs as a type of virtual modality, to augment LiDAR point clouds. The MoDAR modality propagates object information from temporal contexts to a target frame, represented as a set of virtual points, one for each object from a waypoint on a forecasted trajectory. A fused point cloud of both raw sensor points and the virtual points can then be fed to any off-the

-shelf point-cloud based 3D object detector. Evaluated on the Waymo Open Dataset, our method significantly improves prior art detectors by using motion forecasting from extra-long sequences (e.g. 18 seconds), achieving new state of the arts, while not adding much computation overhead.
****

How You Feelin'? Learning Emotions and Mental States in Movie Scenes
Dhruv Srivastava, Aditya Kumar Singh, Makarand Tapaswi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2517-2528
Movie story analysis requires understanding characters' emotions and mental states. Towards this goal, we formulate emotion understanding as predicting a diverse and multi-label set of emotions at the level of a movie scene and for each character. We propose EmoTx, a multimodal Transformer-based architecture that ingests videos, multiple characters, and dialog utterances to make joint predictions. By leveraging annotations from the MovieGraphs dataset, we aim to predict classic emotions (e.g. happy, angry) and other mental states (e.g. honest, helpful). We conduct experiments on the most frequently occurring 10 and 25 labels, and a mapping that clusters 181 labels to 26. Ablation studies and comparison against adapted state-of-the-art emotion recognition approaches shows the effectiveness of EmoTx. Analyzing EmoTx's self-attention scores reveals that expressive emotions often look at character tokens while other mental states rely on video and dialog cues.
****

Dynamic Inference With Grounding Based Vision and Language Models
Burak Uzkent, Amanmeet Garg, Wentao Zhu, Keval Doshi, Jingru Yi, Xiaolong Wang, Mohamed Omar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2624-2633
Transformers have been recently utilized for vision and language tasks successfully. For example, recent image and language models with more than 200M parameters have been proposed to learn visual grounding in the pre-training step and show impressive results on downstream vision and language tasks. On the other hand, there exists a large amount of computational redundancy in these large models which skips their run-time efficiency. To address this problem, we propose dynamic inference for grounding based vision and language models conditioned on the input image-text pair. We first design an approach to dynamically skip multihead self-attention and feed forward network layers across two backbones and multimodal network. Additionally, we propose dynamic token pruning and fusion for two backbones. In particular, we remove redundant tokens at different levels of the backbones and fuse the image tokens with the language tokens in an adaptive manner. To learn policies for dynamic inference, we train agents using reinforcement learning. In this direction, we replace the CNN backbone in a recent grounding-based vision and language model, MDETR, with a vision transformer and call it ViTMDETR. Then, we apply our dynamic inference method to ViTMDETR, called D-ViTDMETR, and perform experiments on image-language tasks. Our results show that we can improve the run-time efficiency of the state-of-the-art models MDETR and GLIP by up to 50% on Referring Expression Comprehension and Segmentation, and VQA with only maximum 0.3% accuracy drop.
****

ALSO: Automotive Lidar Self-Supervision by Occupancy Estimation
Alexandre Boulch, Corentin Sautier, Björn Michele, Gilles Puy, Renaud Marlet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13455-13465
We propose a new self-supervised method for pre-training the backbone of deep perception models operating on point clouds. The core idea is to train the model on a pretext task which is the reconstruction of the surface on which the 3D points are sampled, and to use the underlying latent vectors as input to the perception head. The intuition is that if the network is able to reconstruct the scene surface, given only sparse input points, then it probably also captures some fragments of semantic information, that can be used to boost an actual perception task. This principle has a very simple formulation, which makes it both easy to i

mplement and widely applicable to a large range of 3D sensors and deep networks performing semantic segmentation or object detection. In fact, it supports a single-stream pipeline, as opposed to most contrastive learning approaches, allowing training on limited resources. We conducted extensive experiments on various autonomous driving datasets, involving very different kinds of lidars, for both semantic segmentation and object detection. The results show the effectiveness of our method to learn useful representations without any annotation, compared to existing approaches. The code is available at github.com/valeoai/ALSO

****

Connecting Vision and Language With Video Localized Narratives

Paul Voigtlaender, Soravit Changpinyo, Jordi Pont-Tuset, Radu Soricut, Vittorio Ferrari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2461-2471

We propose Video Localized Narratives, a new form of multimodal video annotations connecting vision and language. In the original Localized Narratives, annotators speak and move their mouse simultaneously on an image, thus grounding each word with a mouse trace segment. However, this is challenging on a video. Our new protocol empowers annotators to tell the story of a video with Localized Narratives, capturing even complex events involving multiple actors interacting with each other and with several passive objects. We annotated 20k videos of the OVIS, UVO, and Oops datasets, totalling 1.7M words. Based on this data, we also construct new benchmarks for the video narrative grounding and video question answering tasks, and provide reference results from strong baseline models. Our annotations are available at https://google.github.io/video-localized-narratives/.

****

Diverse Embedding Expansion Network and Low-Light Cross-Modality Benchmark for Visible-Infrared Person Re-Identification

Yukang Zhang, Hanzi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2153-2162

For the visible-infrared person re-identification (VIReID) task, one of the major challenges is the modality gaps between visible (VIS) and infrared (IR) images. However, the training samples are usually limited, while the modality gaps are too large, which leads that the existing methods cannot effectively mine diverse cross-modality clues. To handle this limitation, we propose a novel augmentation network in the embedding space, called diverse embedding expansion network (DEEN). The proposed DEEN can effectively generate diverse embeddings to learn the informative feature representations and reduce the modality discrepancy between the VIS and IR images. Moreover, the VIReID model may be seriously affected by drastic illumination changes, while all the existing VIReID datasets are captured under sufficient illumination without significant light changes. Thus, we provide a low-light cross-modality (LLCM) dataset, which contains 46,767 bounding boxes of 1,064 identities captured by 9 RGB/IR cameras. Extensive experiments on the SYSU-MM01, RegDB and LLCM datasets show the superiority of the proposed DEEN over several other state-of-the-art methods. The code and dataset are released at: https://github.com/ZYK100/LLCM

****

Model Barrier: A Compact Un-Transferable Isolation Domain for Model Intellectual Property Protection

Lianyu Wang, Meng Wang, Daoqiang Zhang, Huazhu Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20475-20484

As the scientific and technological achievements produced by human intellectual labor and computation cost, model intellectual property (IP) protection, which refers to preventing the usage of the well-trained model on an unauthorized domain, deserves further attention, so as to effectively mobilize the enthusiasm of model owners and creators. To this end, we propose a novel compact un-transferable isolation domain (CUTI-domain), which acts as a model barrier to block illegal transferring from the authorized domain to the unauthorized domain. Specifically, CUTI-domain is investigated to block cross-domain transferring by highlighting private style features of the authorized domain and lead to the failure of rec

ognition on unauthorized domains that contain irrelative private style features. Furthermore, depending on whether the unauthorized domain is known or not, two solutions of using CUTI-domain are provided: target-specified CUTI-domain and target-free CUTI-domain. Comprehensive experimental results on four digit datasets, CIFAR10 & STL10, and VisDA-2017 dataset, demonstrate that our CUTI-domain can be easily implemented with different backbones as a plug-and-play module and provides an efficient solution for model IP protection.
****
Object Detection With Self-Supervised Scene Adaptation
Zekun Zhang, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21589-21599
This paper proposes a novel method to improve the performance of a trained object detector on scenes with fixed camera perspectives based on self-supervised adaptation. Given a specific scene, the trained detector is adapted using pseudo-ground truth labels generated by the detector itself and an object tracker in a cross-teaching manner. When the camera perspective is fixed, our method can utilize the background equivariance by proposing artifact-free object mixup as a means of data augmentation, and utilize accurate background extraction as an additional input modality. We also introduce a large-scale and diverse dataset for the development and evaluation of scene-adaptive object detection. Experiments on this dataset show that our method can improve the average precision of the original detector, outperforming the previous state-of-the-art self-supervised domain adaptive object detection methods by a large margin. Our dataset and code are published at https://github.com/cvlab-stonybrook/scenes100.
****
Visual-Language Prompt Tuning With Knowledge-Guided Context Optimization
Hantao Yao, Rui Zhang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6757-6767
Prompt tuning is an effective way to adapt the pretrained visual-language model (VLM) to the downstream task using task-related textual tokens. Representative CoOp-based works combine the learnable textual tokens with the class tokens to obtain specific textual knowledge. However, the specific textual knowledge has worse generalizable to the unseen classes because it forgets the essential general textual knowledge having a strong generalization ability. To tackle this issue, we introduce a novel Knowledge-guided Context Optimization (KgCoOp) to enhance the generalization ability of the learnable prompt for unseen classes. To remember the essential general knowledge, KgCoOp constructs a regularization term to ensure that the essential general textual knowledge can be embedded into the special textual knowledge generated by the learnable prompt. Especially, KgCoOp minimizes the discrepancy between the textual embeddings generated by learned prompts and the hand-crafted prompts. Finally, adding the KgCoOp upon the contrastive loss can make a discriminative prompt for both seen and unseen tasks. Extensive evaluation of several benchmarks demonstrates that the proposed Knowledge-guided Context Optimization is an efficient method for prompt tuning, i.e., achieves better performance with less training time.
****
Weakly Supervised Video Representation Learning With Unaligned Text for Sequential Videos
Sixun Dong, Huazhang Hu, Dongze Lian, Weixin Luo, Yicheng Qian, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2437-2447
Sequential video understanding, as an emerging video understanding task, has driven lots of researchers' attention because of its goal-oriented nature. This paper studies weakly supervised sequential video understanding where the accurate time-stamp level text-video alignment is not provided. We solve this task by borrowing ideas from CLIP. Specifically, we use a transformer to aggregate frame-level features for video representation and use a pre-trained text encoder to encode the texts corresponding to each action and the whole video, respectively. To model the correspondence between text and video, we propose a multiple granularity loss, where the video-paragraph contrastive loss enforces matching between the

whole video and the complete script, and a fine-grained frame-sentence contrastive loss enforces the matching between each action and its description. As the frame-sentence correspondence is not available, we propose to use the fact that video actions happen sequentially in the temporal domain to generate pseudo frame-sentence correspondence and supervise the network training with the pseudo labels. Extensive experiments on video sequence verification and text-to-video matching show that our method outperforms baselines by a large margin, which validates the effectiveness of our proposed approach. Code is available at https://github.com/svip-lab/WeakSVR.
****
Self-Positioning Point-Based Transformer for Point Cloud Understanding
Jinyoung Park, Sanghyeok Lee, Sihyeon Kim, Yunyang Xiong, Hyunwoo J. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21814-21823
Transformers have shown superior performance on various computer vision tasks with their capabilities to capture long-range dependencies. Despite the success, it is challenging to directly apply Transformers on point clouds due to their quadratic cost in the number of points. In this paper, we present a Self-Positioning point-based Transformer (SPoTr), which is designed to capture both local and global shape contexts with reduced complexity. Specifically, this architecture consists of local self- attention and self-positioning point-based global cross-attention. The self-positioning points, adaptively located based on the input shape, consider both spatial and semantic information with disentangled attention to improve expressive power. With the self-positioning points, we propose a novel global cross-attention mechanism for point clouds, which improves the scalability of global self-attention by allowing the attention module to compute attention weights with only a small set of self-positioning points. Experiments show the effectiveness of SPoTr on three point cloud tasks such as shape classification, part segmentation, and scene segmentation. In particular, our proposed model achieves an accuracy gain of 2.6% over the previous best models on shape classification with ScanObjectNN. We also provide qualitative analyses to demonstrate the interpretability of self-positioning points. The code of SPoTr is available at https://github.com/mlvlab/SPoTr.
****
Bootstrap Your Own Prior: Towards Distribution-Agnostic Novel Class Discovery
Muli Yang, Liancheng Wang, Cheng Deng, Hanwang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3459-3468
Novel Class Discovery (NCD) aims to discover unknown classes without any annotation, by exploiting the transferable knowledge already learned from a base set of known classes. Existing works hold an impractical assumption that the novel class distribution prior is uniform, yet neglect the imbalanced nature of real-world data. In this paper, we relax this assumption by proposing a new challenging task: distribution-agnostic NCD, which allows data drawn from arbitrary unknown class distributions and thus renders existing methods useless or even harmful. We tackle this challenge by proposing a new method, dubbed "Bootstrapping Your Own Prior (BYOP)", which iteratively estimates the class prior based on the model prediction itself. At each iteration, we devise a dynamic temperature technique that better estimates the class prior by encouraging sharper predictions for less-confident samples. Thus, BYOP obtains more accurate pseudo-labels for the novel samples, which are beneficial for the next training iteration. Extensive experiments show that existing methods suffer from imbalanced class distributions, while BYOP outperforms them by clear margins, demonstrating its effectiveness across various distribution scenarios.
****
Learning To Generate Image Embeddings With User-Level Differential Privacy
Zheng Xu, Maxwell Collins, Yuxiao Wang, Liviu Panait, Sewoong Oh, Sean Augenstein, Ting Liu, Florian Schroff, H. Brendan McMahan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7969-7980
Small on-device models have been successfully trained with user-level differenti

al privacy (DP) for next word prediction and image classification tasks in the p
ast. However, existing methods can fail when directly applied to learn embedding
 models using supervised training data with a large class space. To achieve user
-level DP for large image-to-embedding feature extractors, we propose DP-FedEmb,
 a variant of federated learning algorithms with per-user sensitivity control an
d noise addition, to train from user-partitioned data centralized in datacenter.
 DP-FedEmb combines virtual clients, partial aggregation, private local fine-tun
ing, and public pretraining to achieve strong privacy utility trade-offs. We app
ly DP-FedEmb to train image embedding models for faces, landmarks and natural sp
ecies, and demonstrate its superior utility under same privacy budget on benchma
rk datasets DigiFace, GLD and iNaturalist. We further illustrate it is possible
to achieve strong user-level DP guarantees of epsilon < 2 while controlling the
utility drop within 5%, when millions of users can participate in training.
****
Open-Vocabulary Panoptic Segmentation With Text-to-Image Diffusion Models
Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, Shalini De Mell
o; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2023, pp. 2955-2966
We present ODISE: Open-vocabulary DIffusion-based panoptic SEgmentation, which u
nifies pre-trained text-image diffusion and discriminative models to perform ope
n-vocabulary panoptic segmentation. Text-to-image diffusion models have the rema
rkable ability to generate high-quality images with diverse open-vocabulary lang
uage descriptions. This demonstrates that their internal representation space is
 highly correlated with open concepts in the real world. Text-image discriminati
ve models like CLIP, on the other hand, are good at classifying images into open
-vocabulary labels. We leverage the frozen internal representations of both thes
e models to perform panoptic segmentation of any category in the wild. Our appro
ach outperforms the previous state of the art by significant margins on both ope
n-vocabulary panoptic and semantic segmentation tasks. In particular, with COCO
training only, our method achieves 23.4 PQ and 30.0 mIoU on the ADE20K dataset,
with 8.3 PQ and 7.9 mIoU absolute improvement over the previous state of the art
. We open-source our code and models at https://github.com/NVlabs/ODISE.
****
Learning Open-Vocabulary Semantic Segmentation Models From Natural Language Supe
rvision
Jilan Xu, Junlin Hou, Yuejie Zhang, Rui Feng, Yi Wang, Yu Qiao, Weidi Xie; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2023, pp. 2935-2944
In this paper, we consider the problem of open-vocabulary semantic segmentation
(OVS), which aims to segment objects of arbitrary classes instead of pre-defined
, closed-set categories. The main contributions are as follows: First, we propos
e a transformer-based model for OVS, termed as OVSegmentor, which only exploits
web-crawled image-text pairs for pre-training without using any mask annotations
. OVSegmentor assembles the image pixels into a set of learnable group tokens vi
a a slot-attention based binding module, and aligns the group tokens to the corr
esponding caption embedding. Second, we propose two proxy tasks for training, na
mely masked entity completion and cross-image mask consistency. The former aims
to infer all masked entities in the caption given the group tokens, that enables
 the model to learn fine-grained alignment between visual groups and text entiti
es. The latter enforces consistent mask predictions between images that contain
shared entities, which encourages the model to learn visual invariance. Third, w
e construct CC4M dataset for pre-training by filtering CC12M with frequently app
eared entities, which significantly improves training efficiency. Fourth, we per
form zero-shot transfer on three benchmark datasets, PASCAL VOC 2012, PASCAL Con
text, and COCO Object. Our model achieves superior segmentation results over the
 state-of-the-art method by using only 3% data (4M vs 134M) for pre-training. Co
de and pre-trained models will be released for future research.
****
Learning Dynamic Style Kernels for Artistic Style Transfer
Wenju Xu, Chengjiang Long, Yongwei Nie; Proceedings of the IEEE/CVF Conference o

n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10083-10092
Arbitrary style transfer has been demonstrated to be efficient in artistic image generation. Previous methods either globally modulate the content feature ignoring local details, or overly focus on the local structure details leading to style leakage. In contrast to the literature, we propose a new scheme "style kernel" that learns spatially adaptive kernel for per-pixel stylization, where the convolutional kernels are dynamically generated from the global style-content aligned feature and then the learned kernels are applied to modulate the content feature at each spatial position. This new scheme allows flexible both global and local interactions between the content and style features such that the wanted styles can be easily transferred to the content image while at the same time the content structure can be easily preserved. To further enhance the flexibility of our style transfer method, we propose a Style Alignment Encoding (SAE) module complemented with a Content-based Gating Modulation (CGM) module for learning the dynamic style kernels in focusing regions. Extensive experiments strongly demonstrate that our proposed method outperforms state-of-the-art methods and exhibits superior performance in terms of visual quality and efficiency.
****
DeepLSD: Line Segment Detection and Refinement With Deep Image Gradients
Rémi Pautrat, Daniel Barath, Viktor Larsson, Martin R. Oswald, Marc Pollefeys; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17327-17336
Line segments are ubiquitous in our human-made world and are increasingly used in vision tasks. They are complementary to feature points thanks to their spatial extent and the structural information they provide. Traditional line detectors based on the image gradient are extremely fast and accurate, but lack robustness in noisy images and challenging conditions. Their learned counterparts are more repeatable and can handle challenging images, but at the cost of a lower accuracy and a bias towards wireframe lines. We propose to combine traditional and learned approaches to get the best of both worlds: an accurate and robust line detector that can be trained in the wild without ground truth lines. Our new line segment detector, DeepLSD, processes images with a deep network to generate a line attraction field, before converting it to a surrogate image gradient magnitude and angle, which is then fed to any existing handcrafted line detector. Additionally, we propose a new optimization tool to refine line segments based on the attraction field and vanishing points. This refinement improves the accuracy of current deep detectors by a large margin. We demonstrate the performance of our method on low-level line detection metrics, as well as on several downstream tasks using multiple challenging datasets. The source code and models are available at https://github.com/cvg/DeepLSD.
****
OcTr: Octree-Based Transformer for 3D Object Detection
Chao Zhou, Yanan Zhang, Jiaxin Chen, Di Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5166-5175
A key challenge for LiDAR-based 3D object detection is to capture sufficient features from large scale 3D scenes especially for distant or/and occluded objects. Albeit recent efforts made by Transformers with the long sequence modeling capability, they fail to properly balance the accuracy and efficiency, suffering from inadequate receptive fields or coarse-grained holistic correlations. In this paper, we propose an Octree-based Transformer, named OcTr, to address this issue. It first constructs a dynamic octree on the hierarchical feature pyramid through conducting self-attention on the top level and then recursively propagates to the level below restricted by the octants, which captures rich global context in a coarse-to-fine manner while maintaining the computational complexity under control. Furthermore, for enhanced foreground perception, we propose a hybrid positional embedding, composed of the semantic-aware positional embedding and attention mask, to fully exploit semantic and geometry clues. Extensive experiments are conducted on the Waymo Open Dataset and KITTI Dataset, and OcTr reaches newly state-of-the-art results.
****

Chat2Map: Efficient Scene Mapping From Multi-Ego Conversations

Sagnik Majumder, Hao Jiang, Pierre Moulon, Ethan Henderson, Paul Calamia, Kristen Grauman, Vamsi Krishna Ithapu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10554-10564

Can conversational videos captured from multiple egocentric viewpoints reveal the map of a scene in a cost-efficient way? We seek to answer this question by proposing a new problem: efficiently building the map of a previously unseen 3D environment by exploiting shared information in the egocentric audio-visual observations of participants in a natural conversation. Our hypothesis is that as multiple people ("egos") move in a scene and talk among themselves, they receive rich audio-visual cues that can help uncover the unseen areas of the scene. Given the high cost of continuously processing egocentric visual streams, we further explore how to actively coordinate the sampling of visual information, so as to minimize redundancy and reduce power use. To that end, we present an audio-visual deep reinforcement learning approach that works with our shared scene mapper to selectively turn on the camera to efficiently chart out the space. We evaluate the approach using a state-of-the-art audio-visual simulator for 3D scenes as well as real-world video. Our model outperforms previous state-of-the-art mapping methods, and achieves an excellent cost-accuracy tradeoff. Project: https://vision.cs.utexas.edu/projects/chat2map.

****

Learning Distortion Invariant Representation for Image Restoration From a Causality Perspective

Xin Li, Bingchen Li, Xin Jin, Cuiling Lan, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1714-1724

In recent years, we have witnessed the great advancement of Deep neural networks (DNNs) in image restoration. However, a critical limitation is that they cannot generalize well to real-world degradations with different degrees or types. In this paper, we are the first to propose a novel training strategy for image restoration from the causality perspective, to improve the generalization ability of DNNs for unknown degradations. Our method, termed Distortion Invariant representation Learning (DIL), treats each distortion type and degree as one specific confounder, and learns the distortion-invariant representation by eliminating the harmful confounding effect of each degradation. We derive our DIL with the back-door criterion in causality by modeling the interventions of different distortions from the optimization perspective. Particularly, we introduce counterfactual distortion augmentation to simulate the virtual distortion types and degrees as the confounders. Then, we instantiate the intervention of each distortion with a virtual model updating based on corresponding distorted images, and eliminate them from the meta-learning perspective. Extensive experiments demonstrate the generalization capability of our DIL on unseen distortion types and degrees. Our code will be available at https://github.com/lixinustc/Causal-IR-DIL.

****

MOT: Masked Optimal Transport for Partial Domain Adaptation

You-Wei Luo, Chuan-Xian Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3531-3540

As an important methodology to measure distribution discrepancy, optimal transport (OT) has been successfully applied to learn generalizable visual models under changing environments. However, there are still limitations, including strict prior assumption and implicit alignment, for current OT modeling in challenging real-world scenarios like partial domain adaptation, where the learned transport plan may be biased and negative transfer is inevitable. Thus, it is necessary to explore a more feasible OT methodology for real-world applications. In this work, we focus on the rigorous OT modeling for conditional distribution matching and label shift correction. A novel masked OT (MOT) methodology on conditional distributions is proposed by defining a mask operation with label information. Further, a relaxed and reweighting formulation is proposed to improve the robustness of OT in extreme scenarios. We prove the theoretical equivalence between conditional OT and MOT, which implies the well-defined MOT serves as a computation-fri

endly proxy. Extensive experiments validate the effectiveness of theoretical results and proposed model.
****
Executing Your Commands via Motion Diffusion in Latent Space
Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, Gang Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18000-18010
We study a challenging task, conditional human motion generation, which produces plausible human motion sequences according to various conditional inputs, such as action classes or textual descriptors. Since human motions are highly diverse and have a property of quite different distribution from conditional modalities, such as textual descriptors in natural languages, it is hard to learn a probabilistic mapping from the desired conditional modality to the human motion sequences. Besides, the raw motion data from the motion capture system might be redundant in sequences and contain noises; directly modeling the joint distribution over the raw motion sequences and conditional modalities would need a heavy computational overhead and might result in artifacts introduced by the captured noises. To learn a better representation of the various human motion sequences, we first design a powerful Variational AutoEncoder (VAE) and arrive at a representative and low-dimensional latent code for a human motion sequence. Then, instead of using a diffusion model to establish the connections between the raw motion sequences and the conditional inputs, we perform a diffusion process on the motion latent space. Our proposed Motion Latent-based Diffusion model (MLD) could produce vivid motion sequences conforming to the given conditional inputs and substantially reduce the computational overhead in both the training and inference stages. Extensive experiments on various human motion generation tasks demonstrate that our MLD achieves significant improvements over the state-of-the-art methods among extensive human motion generation tasks, with two orders of magnitude faster than previous diffusion models on raw motion sequences.
****
GeoMAE: Masked Geometric Target Prediction for Self-Supervised Point Cloud Pre-Training
Xiaoyu Tian, Haoxi Ran, Yue Wang, Hang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13570-13580
This paper tries to address a fundamental question in point cloud self-supervised learning: what is a good signal we should leverage to learn features from point clouds without annotations? To answer that, we introduce a point cloud representation learning framework, based on geometric feature reconstruction. In contrast to recent papers that directly adopt masked autoencoder (MAE) and only predict original coordinates or occupancy from masked point clouds, our method revisits differences between images and point clouds and identifies three self-supervised learning objectives peculiar to point clouds, namely centroid prediction, normal estimation, and curvature prediction. Combined, these three objectives yield an nontrivial self-supervised learning task and mutually facilitate models to better reason fine-grained geometry of point clouds. Our pipeline is conceptually simple and it consists of two major steps: first, it randomly masks out groups of points, followed by a Transformer-based point cloud encoder; second, a lightweight Transformer decoder predicts centroid, normal, and curvature for points in each voxel. We transfer the pre-trained Transformer encoder to a downstream peception model. On the nuScene Datset, our model achieves 3.38 mAP improvment for object detection, 2.1 mIoU gain for segmentation, and 1.7 AMOTA gain for multi-object tracking. We also conduct experiments on the Waymo Open Dataset and achieve significant performance improvements over baselines as well.
****
Learning Conditional Attributes for Compositional Zero-Shot Learning
Qingsheng Wang, Lingqiao Liu, Chenchen Jing, Hao Chen, Guoqiang Liang, Peng Wang, Chunhua Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11197-11206
Compositional Zero-Shot Learning (CZSL) aims to train models to recognize novel compositional concepts based on learned concepts such as attribute-object combin

ations. One of the challenges is to model attributes interacted with different objects, e.g., the attribute "wet" in "wet apple" and "wet cat" is different. As a solution, we provide analysis and argue that attributes are conditioned on the recognized object and input image and explore learning conditional attribute embeddings by a proposed attribute learning framework containing an attribute hyper learner and an attribute base learner. By encoding conditional attributes, our model enables to generate flexible attribute embeddings for generalization from seen to unseen compositions. Experiments on CZSL benchmarks, including the more challenging C-GQA dataset, demonstrate better performances compared with other state-of-the-art approaches and validate the importance of learning conditional attributes.
****
Complete 3D Human Reconstruction From a Single Incomplete Image
Junying Wang, Jae Shin Yoon, Tuanfeng Y. Wang, Krishna Kumar Singh, Ulrich Neumann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8748-8758
This paper presents a method to reconstruct a complete human geometry and texture from an image of a person with only partial body observed, e.g., a torso. The core challenge arises from the occlusion: there exists no pixel to reconstruct where many existing single-view human reconstruction methods are not designed to handle such invisible parts, leading to missing data in 3D. To address this challenge, we introduce a novel coarse-to-fine human reconstruction framework. For coarse reconstruction, explicit volumetric features are learned to generate a complete human geometry with 3D convolutional neural networks conditioned by a 3D body model and the style features from visible parts. An implicit network combines the learned 3D features with the high-quality surface normals enhanced from multiview to produce fine local details, e.g., high-frequency wrinkles. Finally, we perform progressive texture inpainting to reconstruct a complete appearance of the person in a view-consistent way, which is not possible without the reconstruction of a complete geometry. In experiments, we demonstrate that our method can reconstruct high-quality 3D humans, which is robust to occlusion.
****
PVT-SSD: Single-Stage 3D Object Detector With Point-Voxel Transformer
Honghui Yang, Wenxiao Wang, Minghao Chen, Binbin Lin, Tong He, Hua Chen, Xiaofei He, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13476-13487
Recent Transformer-based 3D object detectors learn point cloud features either from point- or voxel-based representations. However, the former requires time-consuming sampling while the latter introduces quantization errors. In this paper, we present a novel Point-Voxel Transformer for single-stage 3D detection (PVT-SSD) that takes advantage of these two representations. Specifically, we first use voxel-based sparse convolutions for efficient feature encoding. Then, we propose a Point-Voxel Transformer (PVT) module that obtains long-range contexts in a cheap manner from voxels while attaining accurate positions from points. The key to associating the two different representations is our introduced input-dependent Query Initialization module, which could efficiently generate reference points and content queries. Then, PVT adaptively fuses long-range contextual and local geometric information around reference points into content queries. Further, to quickly find the neighboring points of reference points, we design the Virtual Range Image module, which generalizes the native range image to multi-sensor and multi-frame. The experiments on several autonomous driving benchmarks verify the effectiveness and efficiency of the proposed method. Code will be available.
****
Adaptive Human Matting for Dynamic Videos
Chung-Ching Lin, Jiang Wang, Kun Luo, Kevin Lin, Linjie Li, Lijuan Wang, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10229-10238
The most recent efforts in video matting have focused on eliminating trimap dependency since trimap annotations are expensive and trimap-based methods are less adaptable for real-time applications. Despite the latest tripmap-free methods sh

owing promising results, their performance often degrades when dealing with highly diverse and unstructured videos. We address this limitation by introducing Adaptive Matting for Dynamic Videos, termed AdaM, which is a framework designed for simultaneously differentiating foregrounds from backgrounds and capturing alpha matte details of human subjects in the foreground. Two interconnected network designs are employed to achieve this goal: (1) an encoder-decoder network that produces alpha mattes and intermediate masks which are used to guide the transformer in adaptively decoding foregrounds and backgrounds, and (2) a transformer network in which long- and short-term attention combine to retain spatial and temporal contexts, facilitating the decoding of foreground details. We benchmark and study our methods on recently introduced datasets, showing that our model notably improves matting realism and temporal coherence in complex real-world videos and achieves new best-in-class generalizability. Further details and examples are available at https://github.com/microsoft/AdaM.
****
Learning Common Rationale To Improve Self-Supervised Representation for Fine-Grained Visual Recognition Problems
Yangyang Shu, Anton van den Hengel, Lingqiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11392-11401
Self-supervised learning (SSL) strategies have demonstrated remarkable performance in various recognition tasks. However, both our preliminary investigation and recent studies suggest that they may be less effective in learning representations for fine-grained visual recognition (FGVR) since many features helpful for optimizing SSL objectives are not suitable for characterizing the subtle differences in FGVR. To overcome this issue, we propose learning an additional screening mechanism to identify discriminative clues commonly seen across instances and classes, dubbed as common rationales in this paper. Intuitively, common rationales tend to correspond to the discriminative patterns from the key parts of foreground objects. We show that a common rationale detector can be learned by simply exploiting the GradCAM induced from the SSL objective without using any pre-trained object parts or saliency detectors, making it seamlessly to be integrated with the existing SSL process. Specifically, we fit the GradCAM with a branch with limited fitting capacity, which allows the branch to capture the common rationales and discard the less common discriminative patterns. At the test stage, the branch generates a set of spatial weights to selectively aggregate features representing an instance. Extensive experimental results on four visual tasks demonstrate that the proposed method can lead to a significant improvement in different evaluation settings.
****
Reconstructing Animatable Categories From Videos
Gengshan Yang, Chaoyang Wang, N. Dinesh Reddy, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16995-17005
Building animatable 3D models is challenging due to the need for 3D scans, laborious registration, and manual rigging. Recently, differentiable rendering provides a pathway to obtain high-quality 3D models from monocular videos, but these are limited to rigid categories or single instances. We present RAC, a method to build category-level 3D models from monocular videos, disentangling variations over instances and motion over time. Three key ideas are introduced to solve this problem: (1) specializing a category-level skeleton to instances, (2) a method for latent space regularization that encourages shared structure across a category while maintaining instance details, and (3) using 3D background models to disentangle objects from the background. We build 3D models for humans, cats, and dogs given monocular videos. Project page: gengshan-y.github.io/rac-www/
****
UDE: A Unified Driving Engine for Human Motion Generation
Zixiang Zhou, Baoyuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5632-5641
Generating controllable and editable human motion sequences is a key challenge i

n 3D Avatar generation. It has been labor-intensive to generate and animate huma n motion for a long time until learning-based approaches have been developed and applied recently. However, these approaches are still task-specific or modality -specific. In this paper, we propose "UDE", the first unified driving engine tha t enables generating human motion sequences from natural language or audio seque nces (see Fig. 1). Specifically, UDE consists of the following key components: 1) a motion quantization module based on VQVAE that represents continuous motion sequence as discrete latent code, 2) a modality-agnostic transformer encoder tha t learns to map modality-aware driving signals to a joint space, and 3) a unifie d token transformer (GPT-like) network to predict the quantized latent code inde x in an auto-regressive manner. 4) a diffusion motion decoder that takes as inpu t the motion tokens and decodes them into motion sequences with high diversity. We evaluate our method on HumanML3D and AIST++ benchmarks, and the experiment re sults demonstrate our method achieves state-of-the-art performance.
****
High-Fidelity 3D Human Digitization From Single 2K Resolution Images
Sang-Hun Han, Min-Gyu Park, Ju Hong Yoon, Ju-Mi Kang, Young-Jae Park, Hae-Gon Je on; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2023, pp. 12869-12879
High-quality 3D human body reconstruction requires high-fidelity and large-scale training data and appropriate network design that effectively exploits the high -resolution input images. To tackle these problems, we propose a simple yet effe ctive 3D human digitization method called 2K2K, which constructs a large-scale 2 K human dataset and infers 3D human models from 2K resolution images. The propos ed method separately recovers the global shape of a human and its details. The l ow-resolution depth network predicts the global structure from a low-resolution image, and the part-wise image-to-normal network predicts the details of the 3D human body structure. The high-resolution depth network merges the global 3D sha pe and the detailed structures to infer the high-resolution front and back side depth maps. Finally, an off-the-shelf mesh generator reconstructs the full 3D hu man model, which are available at https://github.com/SangHunHan92/2K2K. In addit ion, we also provide 2,050 3D human models, including texture maps, 3D joints, a nd SMPL parameters for research purposes. In experiments, we demonstrate competi tive performance over the recent works on various datasets.
****
Co-Salient Object Detection With Uncertainty-Aware Group Exchange-Masking
Yang Wu, Huihui Song, Bo Liu, Kaihua Zhang, Dong Liu; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19639-19648
The traditional definition of co-salient object detection (CoSOD) task is to seg ment the common salient objects in a group of relevant images. Existing CoSOD mo dels by default adopt the group consensus assumption. This brings about model ro bustness defect under the condition of irrelevant images in the testing image gr oup, which hinders the use of CoSOD models in real-world applications. To addres s this issue, this paper presents a group exchange-masking (GEM) strategy for ro bust CoSOD model learning. With two group of image containing different types of salient object as input, the GEM first selects a set of images from each group by the proposed learning based strategy, then these images are exchanged. The pr oposed feature extraction module considers both the uncertainty caused by the ir relevant images and group consensus in the remaining relevant images. We design a latent variable generator branch which is made of conditional variational auto encoder to generate uncertainly-based global stochastic features. A CoSOD transf ormer branch is devised to capture the correlation-based local features that con tain the group consistency information. At last, the output of two branches are concatenated and fed into a transformer-based decoder, producing robust co-salie ncy prediction. Extensive evaluations on co-saliency detection with and without irrelevant images demonstrate the superiority of our method over a variety of st ate-of-the-art methods.
****
Tangentially Elongated Gaussian Belief Propagation for Event-Based Incremental O

ptical Flow Estimation
Jun Nagata, Yusuke Sekikawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21940-21949
Optical flow estimation is a fundamental functionality in computer vision. An event-based camera, which asynchronously detects sparse intensity changes, is an ideal device for realizing low-latency estimation of the optical flow owing to its low-latency sensing mechanism. An existing method using local plane fitting of events could utilize the sparsity to realize incremental updates for low-latency estimation; however, its output is merely a normal component of the full optical flow. An alternative approach using a frame-based deep neural network could estimate the full flow; however, its intensive non-incremental dense operation prohibits the low-latency estimation. We propose tangentially elongated Gaussian (TEG) belief propagation (BP) that realizes incremental full-flow estimation. We model the probability of full flow as the joint distribution of TEGs from the normal flow measurements, such that the marginal of this distribution with correct prior equals the full flow. We formulate the marginalization using a message-passing based on the BP to realize efficient incremental updates using sparse measurements. In addition to the theoretical justification, we evaluate the effectiveness of the TEGBP in real-world datasets; it outperforms SOTA incremental quasi-full flow method by a large margin. The code will be open-sourced upon acceptance.
****
Extracting Class Activation Maps From Non-Discriminative Features As Well
Zhaozheng Chen, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3135-3144
Extracting class activation maps (CAM) from a classification model often results in poor coverage on foreground objects, i.e., only the discriminative region (e.g., the "head" of "sheep") is recognized and the rest (e.g., the "leg" of "sheep") mistakenly as background. The crux behind is that the weight of the classifier (used to compute CAM) captures only the discriminative features of objects. We tackle this by introducing a new computation method for CAM that explicitly captures non-discriminative features as well, thereby expanding CAM to cover whole objects. Specifically, we omit the last pooling layer of the classification model, and perform clustering on all local features of an object class, where "local" means "at a spatial pixel position". We call the resultant K cluster centers local prototypes - represent local semantics like the "head", "leg", and "body" of "sheep". Given a new image of the class, we compare its unpooled features to every prototype, derive K similarity matrices, and then aggregate them into a heatmap (i.e., our CAM). Our CAM thus captures all local features of the class without discrimination. We evaluate it in the challenging tasks of weakly-supervised semantic segmentation (WSSS), and plug it in multiple state-of-the-art WSSS methods, such as MCTformer and AMN, by simply replacing their original CAM with ours. Our extensive experiments on standard WSSS benchmarks (PASCAL VOC and MS COCO) show the superiority of our method: consistent improvements with little computational overhead.
****
BlendFields: Few-Shot Example-Driven Facial Modeling
Kacper Kania, Stephan J. Garbin, Andrea Tagliasacchi, Virginia Estellers, Kwang Moo Yi, Julien Valentin, Tomasz Trzci■ski, Marek Kowalski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 404-415
Generating faithful visualizations of human faces requires capturing both coarse and fine-level details of the face geometry and appearance. Existing methods are either data-driven, requiring an extensive corpus of data not publicly accessible to the research community, or fail to capture fine details because they rely on geometric face models that cannot represent fine-grained details in texture with a mesh discretization and linear deformation designed to model only a coarse face geometry. We introduce a method that bridges this gap by drawing inspiration from traditional computer graphics techniques. Unseen expressions are modeled by blending appearance from a sparse set of extreme poses. This blending is pe

rformed by measuring local volumetric changes in those expressions and locally r
eproducing their appearance whenever a similar expression is performed at test t
ime. We show that our method generalizes to unseen expressions, adding fine-grai
ned effects on top of smooth volumetric deformations of a face, and demonstrate
how it generalizes beyond faces.
****
Adaptive Sparse Pairwise Loss for Object Re-Identification
Xiao Zhou, Yujie Zhong, Zhen Cheng, Fan Liang, Lin Ma; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19691
-19701
Object re-identification (ReID) aims to find instances with the same identity as
 the given probe from a large gallery. Pairwise losses play an important role in
 training a strong ReID network. Existing pairwise losses densely exploit each i
nstance as an anchor and sample its triplets in a mini-batch. This dense samplin
g mechanism inevitably introduces positive pairs that share few visual similarit
ies, which can be harmful to the training. To address this problem, we propose a
 novel loss paradigm termed Sparse Pairwise (SP) loss that only leverages few ap
propriate pairs for each class in a mini-batch, and empirically demonstrate that
 it is sufficient for the ReID tasks. Based on the proposed loss framework, we p
ropose an adaptive positive mining strategy that can dynamically adapt to divers
e intra-class variations. Extensive experiments show that SP loss and its adapti
ve variant AdaSP loss outperform other pairwise losses, and achieve state-of-the
-art performance across several ReID benchmarks. Code is available at https://gi
thub.com/Astaxanthin/AdaSP.
****
NeFII: Inverse Rendering for Reflectance Decomposition With Near-Field Indirect
Illumination
Haoqian Wu, Zhipeng Hu, Lincheng Li, Yongqiang Zhang, Changjie Fan, Xin Yu; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2023, pp. 4295-4304
Inverse rendering methods aim to estimate geometry, materials and illumination f
rom multi-view RGB images. In order to achieve better decomposition, recent appr
oaches attempt to model indirect illuminations reflected from different material
s via Spherical Gaussians (SG), which, however, tends to blur the high-frequency
 reflection details. In this paper, we propose an end-to-end inverse rendering p
ipeline that decomposes materials and illumination from multi-view images, while
 considering near-field indirect illumination. In a nutshell, we introduce the M
onte Carlo sampling based path tracing and cache the indirect illumination as ne
ural radiance, enabling a physics-faithful and easy-to-optimize inverse renderin
g method. To enhance efficiency and practicality, we leverage SG to represent th
e smooth environment illuminations and apply importance sampling techniques. To
supervise indirect illuminations from unobserved directions, we develop a novel
radiance consistency constraint between implicit neural radiance and path tracin
g results of unobserved rays along with the joint optimization of materials and
illuminations, thus significantly improving the decomposition performance. Exten
sive experiments demonstrate that our method outperforms the state-of-the-art on
 multiple synthetic and real datasets, especially in terms of inter-reflection d
ecomposition.
****
Towards Professional Level Crowd Annotation of Expert Domain Data
Pei Wang, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2023, pp. 3166-3175
Image recognition on expert domains is usually fine-grained and requires expert
labeling, which is costly. This limits dataset sizes and the accuracy of learnin
g systems. To address this challenge, we consider annotating expert data with cr
owdsourcing. This is denoted as PrOfeSsional lEvel cRowd (POSER) annotation. A n
ew approach, based on semi-supervised learning (SSL) and denoted as SSL with hum
an filtering (SSL-HF) is proposed. It is a human-in-the-loop SSL method, where c
rowd-source workers act as filters of pseudo-labels, replacing the unreliable co
nfidence thresholding used by state-of-the-art SSL methods. To enable annotation

by non-experts, classes are specified implicitly, via positive and negative sets of examples and augmented with deliberative explanations, which highlight regions of class ambiguity. In this way, SSL-HF leverages the strong low-shot learning and confidence estimation ability of humans to create an intuitive but effective labeling experience. Experiments show that SSL-HF significantly outperforms various alternative approaches in several benchmarks.
****

Fully Self-Supervised Depth Estimation From Defocus Clue
Haozhe Si, Bin Zhao, Dong Wang, Yunpeng Gao, Mulin Chen, Zhigang Wang, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9140-9149
Depth-from-defocus (DFD), modeling the relationship between depth and defocus pattern in images, has demonstrated promising performance in depth estimation. Recently, several self-supervised works try to overcome the difficulties in acquiring accurate depth ground-truth. However, they depend on the all-in-focus (AIF) images, which cannot be captured in real-world scenarios. Such limitation discourages the applications of DFD methods. To tackle this issue, we propose a completely self-supervised framework that estimates depth purely from a sparse focal stack. We show that our framework circumvents the needs for the depth and AIF image ground-truth, and receives superior predictions, thus closing the gap between the theoretical success of DFD works and their applications in the real world. In particular, we propose (i) a more realistic setting for DFD tasks, where no depth or AIF image ground-truth is available; (ii) a novel self-supervision framework that provides reliable predictions of depth and AIF image under the the challenging setting. The proposed framework uses a neural model to predict the depth and AIF image, and utilizes an optical model to validate and refine the prediction. We verify our framework on three benchmark datasets with rendered focal stacks and real focal stacks. Qualitative and quantitative evaluations show that our method provides a strong baseline for self-supervised DFD tasks. The source code is publicly available at https://github.com/Ehzoahis/DEReD.
****

Semi-Weakly Supervised Object Kinematic Motion Prediction
Gengxin Liu, Qian Sun, Haibin Huang, Chongyang Ma, Yulan Guo, Li Yi, Hui Huang, Ruizhen Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21726-21735
Given a 3D object, kinematic motion prediction aims to identify the mobile parts as well as the corresponding motion parameters. Due to the large variations in both topological structure and geometric details of 3D objects, this remains a challenging task and the lack of large scale labeled data also constrain the performance of deep learning based approaches. In this paper, we tackle the task of object kinematic motion prediction problem in a semi-weakly supervised manner. Our key observations are two-fold. First, although 3D dataset with fully annotated motion labels is limited, there are existing datasets and methods for object part semantic segmentation at large scale. Second, semantic part segmentation and mobile part segmentation is not always consistent but it is possible to detect the mobile parts from the underlying 3D structure. Towards this end, we propose a graph neural network to learn the map between hierarchical part-level segmentation and mobile parts parameters, which are further refined based on geometric alignment. This network can be first trained on PartNet-Mobility dataset with fully labeled mobility information and then applied on PartNet dataset with fine-grained and hierarchical part-level segmentation. The network predictions yield a large scale of 3D objects with pseudo labeled mobility information and can further be used for weakly-supervised learning with pre-existing segmentation. Our experiments show there are significant performance boosts with the augmented data for previous method designed for kinematic motion prediction on 3D partial scans.
****

Learning a Simple Low-Light Image Enhancer From Paired Low-Light Instances
Zhenqi Fu, Yan Yang, Xiaotong Tu, Yue Huang, Xinghao Ding, Kai-Kuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR

), 2023, pp. 22252-22261
Low-light Image Enhancement (LIE) aims at improving contrast and restoring details for images captured in low-light conditions. Most of the previous LIE algorithms adjust illumination using a single input image with several handcrafted priors. Those solutions, however, often fail in revealing image details due to the limited information in a single image and the poor adaptability of handcrafted priors. To this end, we propose PairLIE, an unsupervised approach that learns adaptive priors from low-light image pairs. First, the network is expected to generate the same clean images as the two inputs share the same image content. To achieve this, we impose the network with the Retinex theory and make the two reflectance components consistent. Second, to assist the Retinex decomposition, we propose to remove inappropriate features in the raw image with a simple self-supervised mechanism. Extensive experiments on public datasets show that the proposed PairLIE achieves comparable performance against the state-of-the-art approaches with a simpler network and fewer handcrafted priors. Code is available at: https://github.com/zhenqifu/PairLIE.
****
Deep Stereo Video Inpainting
Zhiliang Wu, Changchang Sun, Hanyu Xuan, Yan Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5693-5702
Stereo video inpainting aims to fill the missing regions on the left and right views of the stereo video with plausible content simultaneously. Compared with the single video inpainting that has achieved promising results using deep convolutional neural networks, inpainting the missing regions of stereo video has not been thoroughly explored. In essence, apart from the spatial and temporal consistency that single video inpainting needs to achieve, another key challenge for stereo video inpainting is to maintain the stereo consistency between left and right views and hence alleviate the 3D fatigue for viewers. In this paper, we propose a novel deep stereo video inpainting network named SVINet, which is the first attempt for stereo video inpainting task utilizing deep convolutional neural networks. SVINet first utilizes a self-supervised flow-guided deformable temporal alignment module to align the features on the left and right view branches, respectively. Then, the aligned features are fed into a shared adaptive feature aggregation module to generate missing contents of their respective branches. Finally, the parallax attention module (PAM) that uses the cross-view information to consider the significant stereo correlation is introduced to fuse the completed features of left and right views. Furthermore, we develop a stereo consistency loss to regularize the trained parameters, so that our model is able to yield high-quality stereo video inpainting results with better stereo consistency. Experimental results demonstrate that our SVINet outperforms state-of-the-art single video inpainting models.
****
Prompting Large Language Models With Answer Heuristics for Knowledge-Based Visual Question Answering
Zhenwei Shao, Zhou Yu, Meng Wang, Jun Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14974-14983
Knowledge-based visual question answering (VQA) requires external knowledge beyond the image to answer the question. Early studies retrieve required knowledge from explicit knowledge bases (KBs), which often introduces irrelevant information to the question, hence restricting the performance of their models. Recent works have sought to use a large language model (i.e., GPT-3) as an implicit knowledge engine to acquire the necessary knowledge for answering. Despite the encouraging results achieved by these methods, we argue that they have not fully activated the capacity of GPT-3 as the provided input information is insufficient. In this paper, we present Prophet---a conceptually simple framework designed to prompt GPT-3 with answer heuristics for knowledge-based VQA. Specifically, we first train a vanilla VQA model on a specific knowledge-based VQA dataset without external knowledge. After that, we extract two types of complementary answer heuristics from the model: answer candidates and answer-aware examples. Finally, the two types of answer heuristics are encoded into the prompts to enable GPT-3 to be

tter comprehend the task thus enhancing its capacity. Prophet significantly outp
erforms all existing state-of-the-art methods on two challenging knowledge-based
 VQA datasets, OK-VQA and A-OKVQA, delivering 61.1% and 55.7% accuracies on thei
r testing sets, respectively.
****
IFSeg: Image-Free Semantic Segmentation via Vision-Language Model
Sukmin Yun, Seong Hyeon Park, Paul Hongsuck Seo, Jinwoo Shin; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp
. 2967-2977
Vision-language (VL) pre-training has recently gained much attention for its tra
nsferability and flexibility in novel concepts (e.g., cross-modality transfer) a
cross various visual tasks. However, VL-driven segmentation has been under-explo
red, and the existing approaches still have the burden of acquiring additional t
raining images or even segmentation annotations to adapt a VL model to downstrea
m segmentation tasks. In this paper, we introduce a novel image-free segmentatio
n task where the goal is to perform semantic segmentation given only a set of th
e target semantic categories, but without any task-specific images and annotatio
ns. To tackle this challenging task, our proposed method, coined IFSeg, generate
s VL-driven artificial image-segmentation pairs and updates a pre-trained VL mod
el to a segmentation task. We construct this artificial training data by creatin
g a 2D map of random semantic categories and another map of their corresponding
word tokens. Given that a pre-trained VL model projects visual and text tokens i
nto a common space where tokens that share the semantics are located closely, th
is artificially generated word map can replace the real image inputs for such a
VL model. Through an extensive set of experiments, our model not only establishe
s an effective baseline for this novel task but also demonstrates strong perform
ances compared to existing methods that rely on stronger supervision, such as ta
sk-specific images and segmentation masks. Code is available at https://github.c
om/alinlab/ifseg.
****
Improving Robustness of Semantic Segmentation to Motion-Blur Using Class-Centric
 Augmentation
Aakanksha, A. N. Rajagopalan; Proceedings of the IEEE/CVF Conference on Computer
 Vision and Pattern Recognition (CVPR), 2023, pp. 10470-10479
Semantic segmentation involves classifying each pixel into one of a pre-defined
set of object/stuff classes. Such a fine-grained detection and localization of o
bjects in the scene is challenging by itself. The complexity increases manifold
in the presence of blur. With cameras becoming increasingly light-weight and com
pact, blur caused by motion during capture time has become unavoidable. Most res
earch has focused on improving segmentation performance for sharp clean images a
nd the few works that deal with degradations, consider motion-blur as one of man
y generic degradations. In this work, we focus exclusively on motion-blur and at
tempt to achieve robustness for semantic segmentation in its presence. Based on
the observation that segmentation annotations can be used to generate synthetic
space-variant blur, we propose a Class-Centric Motion-Blur Augmentation (CCMBA)
strategy. Our approach involves randomly selecting a subset of semantic classes
present in the image and using the segmentation map annotations to blur only the
 corresponding regions. This enables the network to simultaneously learn semanti
c segmentation for clean images, images with egomotion blur, as well as images w
ith dynamic scene blur. We demonstrate the effectiveness of our approach for bot
h CNN and Vision Transformer-based semantic segmentation networks on PASCAL VOC
and Cityscapes datasets. We also illustrate the improved generalizability of our
 method to complex real-world blur by evaluating on the commonly used deblurring
 datasets GoPro and REDS.
****
Progressive Open Space Expansion for Open-Set Model Attribution
Tianyun Yang, Danding Wang, Fan Tang, Xinying Zhao, Juan Cao, Sheng Tang; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 15856-15865
Despite the remarkable progress in generative technology, the Janus-faced issues

 of intellectual property protection and malicious content supervision have aris
en. Efforts have been paid to manage synthetic images by attributing them to a s
et of potential source models. However, the closed-set classification setting li
mits the application in real-world scenarios for handling contents generated by
arbitrary models. In this study, we focus on a challenging task, namely Open-Set
 Model Attribution (OSMA), to simultaneously attribute images to known models an
d identify those from unknown ones. Compared to existing open-set recognition (O
SR) tasks focusing on semantic novelty, OSMA is more challenging as the distinct
ion between images from known and unknown models may only lie in visually imperc
eptible traces. To this end, we propose a Progressive Open Space Expansion (POSE
) solution, which simulates open-set samples that maintain the same semantics as
 closed-set samples but embedded with different imperceptible traces. Guided by
a diversity constraint, the open space is simulated progressively by a set of li
ghtweight augmentation models. We consider three real-world scenarios and constr
uct an OSMA benchmark dataset, including unknown models trained with different r
andom seeds, architectures, and datasets from known ones. Extensive experiments
on the dataset demonstrate POSE is superior to both existing model attribution m
ethods and off-the-shelf OSR methods.
****
Backdoor Cleansing With Unlabeled Data
Lu Pang, Tao Sun, Haibin Ling, Chao Chen; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12218-12227
Due to the increasing computational demand of Deep Neural Networks (DNNs), compa
nies and organizations have begun to outsource the training process. However, th
e externally trained DNNs can potentially be backdoor attacked. It is crucial to
 defend against such attacks, i.e, to postprocess a suspicious model so that its
 backdoor behavior is mitigated while its normal prediction power on clean input
s remain uncompromised. To remove the abnormal backdoor behavior, existing metho
ds mostly rely on additional labeled clean samples. However, such requirement ma
y be unrealistic as the training data are often unavailable to end users. In thi
s paper, we investigate the possibility of circumventing such barrier. We propos
e a novel defense method that does not require training labels. Through a carefu
lly designed layer-wise weight re-initialization and knowledge distillation, our
 method can effectively cleanse backdoor behaviors of a suspicious network with
 negligible compromise in its normal behavior. In experiments, we show that our
 method, trained without labels, is on-par with state-of-the-art defense methods
 trained using labels. We also observe promising defense results even on out-of-
distribution data. This makes our method very practical. Code is available at: h
ttps://github.com/luluppang/BCU.
****
Is BERT Blind? Exploring the Effect of Vision-and-Language Pretraining on Visual
 Language Understanding
Morris Alper, Michael Fiman, Hadar Averbuch-Elor; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6778-6788
Most humans use visual imagination to understand and reason about language, but
models such as BERT reason about language using knowledge acquired during text-o
nly pretraining. In this work, we investigate whether vision-and-language pretra
ining can improve performance on text-only tasks that involve implicit visual re
asoning, focusing primarily on zero-shot probing methods. We propose a suite of
visual language understanding (VLU) tasks for probing the visual reasoning abili
ties of text encoder models, as well as various non-visual natural language unde
rstanding (NLU) tasks for comparison. We also contribute a novel zero-shot knowl
edge probing method, Stroop probing, for applying models such as CLIP to text-on
ly tasks without needing a prediction head such as the masked language modelling
 head of models like BERT. We show that SOTA multimodally trained text encoders
outperform unimodally trained text encoders on the VLU tasks while being underpe
rformed by them on the NLU tasks, lending new context to previously mixed result
s regarding the NLU capabilities of multimodal models. We conclude that exposure
 to images during pretraining affords inherent visual reasoning knowledge that i
s reflected in language-only tasks that require implicit visual reasoning. Our f

indings bear importance in the broader context of multimodal learning, providing principled guidelines for the choice of text encoders used in such contexts.
****

PivoTAL: Prior-Driven Supervision for Weakly-Supervised Temporal Action Localization

Mamshad Nayeem Rizve, Gaurav Mittal, Ye Yu, Matthew Hall, Sandra Sajeev, Mubarak Shah, Mei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22992-23002

Weakly-supervised Temporal Action Localization (WTAL) attempts to localize the actions in untrimmed videos using only video-level supervision. Most recent works approach WTAL from a localization-by-classification perspective where these methods try to classify each video frame followed by a manually-designed post-processing pipeline to aggregate these per-frame action predictions into action snippets. Due to this perspective, the model lacks any explicit understanding of action boundaries and tends to focus only on the most discriminative parts of the video resulting in incomplete action localization. To address this, we present PivoTAL, Prior-driven Supervision for Weakly-supervised Temporal Action Localization, to approach WTAL from a localization-by-localization perspective by learning to localize the action snippets directly. To this end, PivoTAL leverages the underlying spatio-temporal regularities in videos in the form of action-specific scene prior, action snippet generation prior, and learnable Gaussian prior to supervise the localization-based training. PivoTAL shows significant improvement (of at least 3% avg mAP) over all existing methods on the benchmark datasets, THUMOS-14 and ActivitNet-v1.3.
****

Harmonious Feature Learning for Interactive Hand-Object Pose Estimation

Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, Shaoli Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12989-12998

Joint hand and object pose estimation from a single image is extremely challenging as serious occlusion often occurs when the hand and object interact. Existing approaches typically first extract coarse hand and object features from a single backbone, then further enhance them with reference to each other via interaction modules. However, these works usually ignore that the hand and object are competitive in feature learning, since the backbone takes both of them as foreground and they are usually mutually occluded. In this paper, we propose a novel Harmonious Feature Learning Network (HFL-Net). HFL-Net introduces a new framework that combines the advantages of single- and double-stream backbones: it shares the parameters of the low- and high-level convolutional layers of a common ResNet-50 model for the hand and object, leaving the middle-level layers unshared. This strategy enables the hand and the object to be extracted as the sole targets by the middle-level layers, avoiding their competition in feature learning. The shared high-level layers also force their features to be harmonious, thereby facilitating their mutual feature enhancement. In particular, we propose to enhance the feature of the hand via concatenation with the feature in the same location from the object stream. A subsequent self-attention layer is adopted to deeply fuse the concatenated feature. Experimental results show that our proposed approach consistently outperforms state-of-the-art methods on the popular HO3D and Dex-YCB databases. Notably, the performance of our model on hand pose estimation even surpasses that of existing works that only perform the single-hand pose estimation task. Code is available at https://github.com/lzfff12/HFL-Net.
****

3D GAN Inversion With Facial Symmetry Prior

Fei Yin, Yong Zhang, Xuan Wang, Tengfei Wang, Xiaoyu Li, Yuan Gong, Yanbo Fan, Xiaodong Cun, Ying Shan, Cengiz Oztireli, Yujiu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 342-351

Recently, a surge of high-quality 3D-aware GANs have been proposed, which leverage the generative power of neural rendering. It is natural to associate 3D GANs with GAN inversion methods to project a real image into the generator's latent space, allowing free-view consistent synthesis and editing, referred as 3D GAN in

version. Although with the facial prior preserved in pre-trained 3D GANs, recons
tructing a 3D portrait with only one monocular image is still an ill-pose proble
m. The straightforward application of 2D GAN inversion methods focuses on textur
e similarity only while ignoring the correctness of 3D geometry shapes. It may r
aise geometry collapse effects, especially when reconstructing a side face under
 an extreme pose. Besides, the synthetic results in novel views are prone to be
blurry. In this work, we propose a novel method to promote 3D GAN inversion by i
ntroducing facial symmetry prior. We design a pipeline and constraints to make f
ull use of the pseudo auxiliary view obtained via image flipping, which helps ob
tain a view-consistent and well-structured geometry shape during the inversion p
rocess. To enhance texture fidelity in unobserved viewpoints, pseudo labels from
 depth-guided 3D warping can provide extra supervision. We design constraints ai
med at filtering out conflict areas for optimization in asymmetric situations. C
omprehensive quantitative and qualitative evaluations on image reconstruction an
d editing demonstrate the superiority of our method.
****
CLOTH4D: A Dataset for Clothed Human Reconstruction
Xingxing Zou, Xintong Han, Waikeung Wong; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12847-12857
Clothed human reconstruction is the cornerstone for creating the virtual world.
To a great extent, the quality of recovered avatars decides whether the Metavers
e is a passing fad. In this work, we introduce CLOTH4D, a clothed human dataset
containing 1,000 subjects with varied appearances, 1,000 3D outfits, and over 10
0,000 clothed meshes with paired unclothed humans, to fill the gap in large-scal
e and high-quality 4D clothing data. It enjoys appealing characteristics: 1) Acc
urate and detailed clothing textured meshes---all clothing items are manually cr
eated and then simulated in professional software, strictly following the genera
l standard in fashion design. 2) Separated textured clothing and under-clothing
body meshes, closer to the physical world than single-layer raw scans. 3) Clothe
d human motion sequences simulated given a set of 289 actions, covering fundamen
tal and complicated dynamics. Upon CLOTH4D, we novelly designed a series of temp
orally-aware metrics to evaluate the temporal stability of the generated 3D huma
n meshes, which has been overlooked previously. Moreover, by assessing and retra
ining current state-of-the-art clothed human reconstruction methods, we reveal i
nsights, present improved performance, and propose potential future research dir
ections, confirming our dataset's advancement. The dataset is available at www.g
ithub.com/AemikaChow/AiDLab-fAshIon-Data
****
SDFusion: Multimodal 3D Shape Completion, Reconstruction, and Generation
Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G. Schwing, Liang-Yan G
ui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2023, pp. 4456-4465
In this work, we present a novel framework built to simplify 3D asset generation
 for amateur users. To enable interactive generation, our method supports a vari
ety of input modalities that can be easily provided by a human, including images
, texts, partially observed shapes and combinations of these, further allowing f
or adjusting the strength of each input. At the core of our approach is an encod
er-decoder, compressing 3D shapes into a compact latent representation, upon whi
ch a diffusion model is learned. To enable a variety of multi-modal inputs, we e
mploy task-specific encoders with dropout followed by a cross-attention mechanis
m. Due to its flexibility, our model naturally supports a variety of tasks outpe
rforming prior works on shape completion, image-based 3D reconstruction, and tex
t-to-3D. Most interestingly, our model can combine all these tasks into one swis
s-army-knife tool, enabling the user to perform shape generation using incomplet
e shapes, images, and textual descriptions at the same time, providing the relat
ive weights for each input and facilitating interactivity. Despite our approach
being shape-only, we further show an efficient method to texture the generated u
sing large-scale text-to-image models.
****
SMAE: Few-Shot Learning for HDR Deghosting With Saturation-Aware Masked Autoenco

ders
Qingsen Yan, Song Zhang, Weiye Chen, Hao Tang, Yu Zhu, Jinqiu Sun, Luc Van Gool, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5775-5784
Generating a high-quality High Dynamic Range (HDR) image from dynamic scenes has recently been extensively studied by exploiting Deep Neural Networks (DNNs). Most DNNs-based methods require a large amount of training data with ground truth, requiring tedious and time-consuming work. Few-shot HDR imaging aims to generate satisfactory images with limited data. However, it is difficult for modern DNNs to avoid overfitting when trained on only a few images. In this work, we propose a novel semi-supervised approach to realize few-shot HDR imaging via two stages of training, called SSHDR. Unlikely previous methods, directly recovering content and removing ghosts simultaneously, which is hard to achieve optimum, we first generate content of saturated regions with a self-supervised mechanism and then address ghosts via an iterative semi-supervised learning framework. Concretely, considering that saturated regions can be regarded as masking Low Dynamic Range (LDR) input regions, we design a Saturated Mask AutoEncoder (SMAE) to learn a robust feature representation and reconstruct a non-saturated HDR image. We also propose an adaptive pseudo-label selection strategy to pick high-quality HDR pseudo-labels in the second stage to avoid the effect of mislabeled samples. Experiments demonstrate that SSHDR outperforms state-of-the-art methods quantitatively and qualitatively within and across different datasets, achieving appealing HDR visualization with few labeled samples.
****
Improving Generalization With Domain Convex Game
Fangrui Lv, Jian Liang, Shuang Li, Jinming Zhang, Di Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24315-24324
Domain generalization (DG) tends to alleviate the poor generalization capability of deep neural networks by learning model with multiple source domains. A classical solution to DG is domain augmentation, the common belief of which is that diversifying source domains will be conducive to the out-of-distribution generalization. However, these claims are understood intuitively, rather than mathematically. Our explorations empirically reveal that the correlation between model generalization and the diversity of domains may be not strictly positive, which limits the effectiveness of domain augmentation. This work therefore aim to guarantee and further enhance the validity of this strand. To this end, we propose a new perspective on DG that recasts it as a convex game between domains. We first encourage each diversified domain to enhance model generalization by elaborately designing a regularization term based on supermodularity. Meanwhile, a sample filter is constructed to eliminate low-quality samples, thereby avoiding the impact of potentially harmful information. Our framework presents a new avenue for the formal analysis of DG, heuristic analysis and extensive experiments demonstrate the rationality and effectiveness.
****
Learning To Render Novel Views From Wide-Baseline Stereo Pairs
Yilun Du, Cameron Smith, Ayush Tewari, Vincent Sitzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4970-4980
We introduce a method for novel view synthesis given only a single wide-baseline stereo image pair. In this challenging regime, 3D scene points are regularly observed only once, requiring prior-based reconstruction of scene geometry and appearance. We find that existing approaches to novel view synthesis from sparse observations fail due to recovering incorrect 3D geometry and the high cost of differentiable rendering that precludes their scaling to large-scale training. We take a step towards resolving these shortcomings by formulating a multi-view transformer encoder, proposing an efficient, image-space epipolar line sampling scheme to assemble image features for a target ray, and a lightweight cross-attention-based renderer. Our contributions enable training of our method on a large-scale real-world dataset of indoor and outdoor scenes. In several ablation studies,

we demonstrate that our contributions enable learning of powerful multi-view geometry priors while reducing both rendering time and memory footprint. We conduct extensive comparisons on held-out test scenes across two real-world datasets, significantly outperforming prior work on novel view synthesis from sparse image observations and achieving multi-view-consistent novel view synthesis.
****

TryOnDiffusion: A Tale of Two UNets
Luyang Zhu, Dawei Yang, Tyler Zhu, Fitsum Reda, William Chan, Chitwan Saharia, Mohammad Norouzi, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4606-4615
Given two images depicting a person and a garment worn by another person, our goal is to generate a visualization of how the garment might look on the input person. A key challenge is to synthesize a photorealistic detail-preserving visualization of the garment, while warping the garment to accommodate a significant body pose and shape change across the subjects. Previous methods either focus on garment detail preservation without effective pose and shape variation, or allow try-on with the desired shape and pose but lack garment details. In this paper, we propose a diffusion-based architecture that unifies two UNets (referred to as Parallel-UNet), which allows us to preserve garment details and warp the garment for significant pose and body change in a single network. The key ideas behind Parallel-UNet include: 1) garment is warped implicitly via a cross attention mechanism, 2) garment warp and person blend happen as part of a unified process as opposed to a sequence of two separate tasks. Experimental results indicate that TryOnDiffusion achieves state-of-the-art performance both qualitatively and quantitatively.
****

Fair Scratch Tickets: Finding Fair Sparse Networks Without Weight Training
Pengwei Tang, Wei Yao, Zhicong Li, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24406-24416
Recent studies suggest that computer vision models come at the risk of compromising fairness. There are extensive works to alleviate unfairness in computer vision using pre-processing, in-processing, and post-processing methods. In this paper, we lead a novel fairness-aware learning paradigm for in-processing methods through the lens of the lottery ticket hypothesis (LTH) in the context of computer vision fairness. We randomly initialize a dense neural network and find appropriate binary masks for the weights to obtain fair sparse subnetworks without any weight training. Interestingly, to the best of our knowledge, we are the first to discover that such sparse subnetworks with inborn fairness exist in randomly initialized networks, achieving an accuracy-fairness trade-off comparable to that of dense neural networks trained with existing fairness-aware in-processing approaches. We term these fair subnetworks as Fair Scratch Tickets (FSTs). We also theoretically provide fairness and accuracy guarantees for them. In our experiments, we investigate the existence of FSTs on various datasets, target attributes, random initialization methods, sparsity patterns, and fairness surrogates. We also find that FSTs can transfer across datasets and investigate other properties of FSTs.
****

Generative Bias for Robust Visual Question Answering
Jae Won Cho, Dong-Jin Kim, Hyeonggon Ryu, In So Kweon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11681-11690
The task of Visual Question Answering (VQA) is known to be plagued by the issue of VQA models exploiting biases within the dataset to make its final prediction. Various previous ensemble based debiasing methods have been proposed where an additional model is purposefully trained to be biased in order to train a robust target model. However, these methods compute the bias for a model simply from the label statistics of the training data or from single modal branches. In this work, in order to better learn the bias a target VQA model suffers from, we propose a generative method to train the bias model directly from the target model, called GenB. In particular, GenB employs a generative network to learn the bias i

n the target model through a combination of the adversarial objective and knowledge distillation. We then debias our target model with GenB as a bias model, and show through extensive experiments the effects of our method on various VQA bias datasets including VQA-CP2, VQA-CP1, GQA-OOD, and VQA-CE, and show state-of-the-art results with the LXMERT architecture on VQA-CP2.
****
Data-Free Sketch-Based Image Retrieval
Abhra Chaudhuri, Ayan Kumar Bhunia, Yi-Zhe Song, Anjan Dutta; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12084-12093
Rising concerns about privacy and anonymity preservation of deep learning models have facilitated research in data-free learning. Primarily based on data-free knowledge distillation, models developed in this area so far have only been able to operate in a single modality, performing the same kind of task as that of the teacher. For the first time, we propose Data-Free Sketch-Based Image Retrieval (DF-SBIR), a cross-modal data-free learning setting, where teachers trained for classification in a single modality have to be leveraged by students to learn a cross-modal metric-space for retrieval. The widespread availability of pre-trained classification models, along with the difficulty in acquiring paired photo-sketch datasets for SBIR justify the practicality of this setting. We present a methodology for DF-SBIR, which can leverage knowledge from models independently trained to perform classification on photos and sketches. We evaluate our model on the Sketchy, TU-Berlin, and QuickDraw benchmarks, designing a variety of baselines based on existing data-free learning literature, and observe that our method surpasses all of them by significant margins. Our method also achieves mAPs competitive with data-dependent approaches, all the while requiring no training data. Implementation is available at https://github.com/abhrac/data-free-sbir.
****
Multi-Object Manipulation via Object-Centric Neural Scattering Functions
Stephen Tian, Yancheng Cai, Hong-Xing Yu, Sergey Zakharov, Katherine Liu, Adrien Gaidon, Yunzhu Li, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9021-9031
Learned visual dynamics models have proven effective for robotic manipulation tasks. Yet, it remains unclear how best to represent scenes involving multi-object interactions. Current methods decompose a scene into discrete objects, yet they struggle with precise modeling and manipulation amid challenging lighting conditions since they only encode appearance tied with specific illuminations. In this work, we propose using object-centric neural scattering functions (OSFs) as object representations in a model-predictive control framework. OSFs model per-object light transport, enabling compositional scene re-rendering under object rearrangement and varying lighting conditions. By combining this approach with inverse parameter estimation and graph-based neural dynamics models, we demonstrate improved model-predictive control performance and generalization in compositional multi-object environments, even in previously unseen scenarios and harsh lighting conditions.
****
The Wisdom of Crowds: Temporal Progressive Attention for Early Action Prediction
Alexandros Stergiou, Dima Damen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14709-14719
Early action prediction deals with inferring the ongoing action from partially-observed videos, typically at the outset of the video. We propose a bottleneck-based attention model that captures the evolution of the action, through progressive sampling over fine-to-coarse scales. Our proposed Temporal Progressive (TemPr) model is composed of multiple attention towers, one for each scale. The predicted action label is based on the collective agreement considering confidences of these towers. Extensive experiments over four video datasets showcase state-of-the-art performance on the task of Early Action Prediction across a range of encoder architectures. We demonstrate the effectiveness and consistency of TemPr through detailed ablations.
****

Invertible Neural Skinning
Yash Kant, Aliaksandr Siarohin, Riza Alp Guler, Menglei Chai, Jian Ren, Sergey Tulyakov, Igor Gilitschenski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8715-8725
Building animatable and editable models of clothed humans from raw 3D scans and poses is a challenging problem. Existing reposing methods suffer from the limited expressiveness of Linear Blend Skinning (LBS), require costly mesh extraction to generate each new pose, and typically do not preserve surface correspondences across different poses. In this work, we introduce Invertible Neural Skinning (INS) to address these shortcomings. To maintain correspondences, we propose a Pose-conditioned Invertible Network (PIN) architecture, which extends the LBS process by learning additional pose-varying deformations. Next, we combine PIN with a differentiable LBS module to build an expressive and end-to-end Invertible Neural Skinning (INS) pipeline. We demonstrate the strong performance of our method by outperforming the state-of-the-art reposing techniques on clothed humans and preserving surface correspondences, while being an order of magnitude faster. We also perform an ablation study, which shows the usefulness of our pose-conditioning formulation, and our qualitative results display that INS can rectify artefacts introduced by LBS well.
****
Weakly Supervised Semantic Segmentation via Adversarial Learning of Classifier and Reconstructor
Hyeokjun Kweon, Sung-Hoon Yoon, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11329-11339
In Weakly Supervised Semantic Segmentation (WSSS), Class Activation Maps (CAMs) usually 1) do not cover the whole object and 2) be activated on irrelevant regions. To address the issues, we propose a novel WSSS framework via adversarial learning of a classifier and an image reconstructor. When an image is perfectly decomposed into class-wise segments, information (i.e., color or texture) of a single segment could not be inferred from the other segments. Therefore, inferability between the segments can represent the preciseness of segmentation. We quantify the inferability as a reconstruction quality of one segment from the other segments. If one segment could be reconstructed from the others, then the segment would be imprecise. To bring this idea into WSSS, we simultaneously train two models: a classifier generating CAMs that decompose an image into segments and a reconstructor that measures the inferability between the segments. As in GANs, while being alternatively trained in an adversarial manner, two networks provide positive feedback to each other. We verify the superiority of the proposed framework with extensive ablation studies. Our method achieves new state-of-the-art performances on both PASCAL VOC 2012 and MS COCO 2014. The code is available at https://github.com/sangrockEG/ACR.
****
Intrinsic Physical Concepts Discovery With Object-Centric Predictive Models
Qu Tang, Xiangyu Zhu, Zhen Lei, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23252-23261
The ability to discover abstract physical concepts and understand how they work in the world through observing lies at the core of human intelligence. The acquisition of this ability is based on compositionally perceiving the environment in terms of objects and relations in an unsupervised manner. Recent approaches learn object-centric representations and capture visually observable concepts of objects, e.g., shape, size, and location. In this paper, we take a step forward and try to discover and represent intrinsic physical concepts such as mass and charge. We introduce the PHYsical Concepts Inference NEtwork (PHYCINE), a system that infers physical concepts in different abstract levels without supervision. The key insights underlining PHYCINE are two-fold, commonsense knowledge emerges with prediction, and physical concepts of different abstract levels should be reasoned in a bottom-to-up fashion. Empirical evaluation demonstrates that variables inferred by our system work in accordance with the properties of the corresponding physical concepts. We also show that object representations containing the discovered physical concepts variables could help achieve better performance in

causal reasoning tasks, i.e., COMPHY.
****
Distilling Cross-Temporal Contexts for Continuous Sign Language Recognition
Leming Guo, Wanli Xue, Qing Guo, Bo Liu, Kaihua Zhang, Tiantian Yuan, Shengyong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10771-10780
Continuous sign language recognition (CSLR) aims to recognize glosses in a sign language video. State-of-the-art methods typically have two modules, a spatial perception module and a temporal aggregation module, which are jointly learned end-to-end. Existing results in [9,20,25,36] have indicated that, as the frontal component of the overall model, the spatial perception module used for spatial feature extraction tends to be insufficiently trained. In this paper, we first conduct empirical studies and show that a shallow temporal aggregation module allows more thorough training of the spatial perception module. However, a shallow temporal aggregation module cannot well capture both local and global temporal context information in sign language. To address this dilemma, we propose a cross-temporal context aggregation (CTCA) model. Specifically, we build a dual-path network that contains two branches for perceptions of local temporal context and global temporal context. We further design a cross-context knowledge distillation learning objective to aggregate the two types of context and the linguistic prior. The knowledge distillation enables the resultant one-branch temporal aggregation module to perceive local-global temporal and semantic context. This shallow temporal perception module structure facilitates spatial perception module learning. Extensive experiments on challenging CSLR benchmarks demonstrate that our method outperforms all state-of-the-art methods.
****
Automatic High Resolution Wire Segmentation and Removal
Mang Tik Chiu, Xuaner Zhang, Zijun Wei, Yuqian Zhou, Eli Shechtman, Connelly Barnes, Zhe Lin, Florian Kainz, Sohrab Amirghodsi, Humphrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2183-2192
Wires and powerlines are common visual distractions that often undermine the aesthetics of photographs. The manual process of precisely segmenting and removing them is extremely tedious and may take up to hours, especially on high-resolution photos where wires may span the entire space. In this paper, we present an automatic wire clean-up system that eases the process of wire segmentation and removal/inpainting to within a few seconds. We observe several unique challenges: wires are thin, lengthy, and sparse. These are rare properties of subjects that common segmentation tasks cannot handle, especially in high-resolution images. We thus propose a two-stage method that leverages both global and local context to accurately segment wires in high-resolution images efficiently, and a tile-based inpainting strategy to remove the wires given our predicted segmentation masks. We also introduce the first wire segmentation benchmark dataset, WireSegHR. Finally, we demonstrate quantitatively and qualitatively that our wire clean-up system enables fully automated wire removal for great generalization to various wire appearances.
****
The Resource Problem of Using Linear Layer Leakage Attack in Federated Learning
Joshua C. Zhao, Ahmed Roushdy Elkordy, Atul Sharma, Yahya H. Ezzeldin, Salman Avestimehr, Saurabh Bagchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3974-3983
Secure aggregation promises a heightened level of privacy in federated learning, maintaining that a server only has access to a decrypted aggregate update. Within this setting, linear layer leakage methods are the only data reconstruction attacks able to scale and achieve a high leakage rate regardless of the number of clients or batch size. This is done through increasing the size of an injected fully-connected (FC) layer. We show that this results in a resource overhead which grows larger with an increasing number of clients. We show that this resource overhead is caused by an incorrect perspective in all prior work that treats an attack on an aggregate update in the same way as an individual update with a la

rger batch size. Instead, by attacking the update from the perspective that aggregation is combining multiple individual updates, this allows the application of sparsity to alleviate resource overhead. We show that the use of sparsity can decrease the model size overhead by over 327x and the computation time by 3.34x compared to SOTA while maintaining equivalent total leakage rate, 77% even with 1000 clients in aggregation.
****
Unsupervised Deep Probabilistic Approach for Partial Point Cloud Registration
Guofeng Mei, Hao Tang, Xiaoshui Huang, Weijie Wang, Juan Liu, Jian Zhang, Luc Van Gool, Qiang Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13611-13620
Deep point cloud registration methods face challenges to partial overlaps and rely on labeled data. To address these issues, we propose UDPReg, an unsupervised deep probabilistic registration framework for point clouds with partial overlaps. Specifically, we first adopt a network to learn posterior probability distributions of Gaussian mixture models (GMMs) from point clouds. To handle partial point cloud registration, we apply the Sinkhorn algorithm to predict the distribution-level correspondences under the constraint of the mixing weights of GMMs. To enable unsupervised learning, we design three distribution consistency-based losses: self-consistency, cross-consistency, and local contrastive. The self-consistency loss is formulated by encouraging GMMs in Euclidean and feature spaces to share identical posterior distributions. The cross-consistency loss derives from the fact that the points of two partially overlapping point clouds belonging to the same clusters share the cluster centroids. The cross-consistency loss allows the network to flexibly learn a transformation-invariant posterior distribution of two aligned point clouds. The local contrastive loss facilitates the network to extract discriminative local features. Our UDPReg achieves competitive performance on the 3DMatch/3DLoMatch and ModelNet/ModelLoNet benchmarks.
****
Towards Generalisable Video Moment Retrieval: Visual-Dynamic Injection to Image-Text Pre-Training
Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23045-23055
The correlation between the vision and text is essential for video moment retrieval (VMR), however, existing methods heavily rely on separate pre-training feature extractors for visual and textual understanding. Without sufficient temporal boundary annotations, it is non-trivial to learn universal video-text alignments. In this work, we explore multi-modal correlations derived from large-scale image-text data to facilitate generalisable VMR. To address the limitations of image-text pre-training models on capturing the video changes, we propose a generic method, referred to as Visual-Dynamic Injection (VDI), to empower the model's understanding of video moments. Whilst existing VMR methods are focusing on building temporal-aware video features, being aware of the text descriptions about the temporal changes is also critical but originally overlooked in pre-training by matching static images with sentences. Therefore, we extract visual context and spatial dynamic information from video frames and explicitly enforce their alignments with the phrases describing video changes (e.g. verb). By doing so, the potentially relevant visual and motion patterns in videos are encoded in the corresponding text embeddings (injected) so to enable more accurate video-text alignments. We conduct extensive experiments on two VMR benchmark datasets (Charades-STA and ActivityNet-Captions) and achieve state-of-the-art performances. Especially, VDI yields notable advantages when being tested on the out-of-distribution splits where the testing samples involve novel scenes and vocabulary.
****
Learning Adaptive Dense Event Stereo From the Image Domain
Hoonhee Cho, Jegyeong Cho, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17797-17807
Recently, event-based stereo matching has been studied due to its robustness in poor light conditions. However, existing event-based stereo networks suffer seve

re performance degradation when domains shift. Unsupervised domain adaptation (U DA) aims at resolving this problem without using the target domain ground-truth. However, traditional UDA still needs the input event data with ground-truth in the source domain, which is more challenging and costly to obtain than image dat a. To tackle this issue, we propose a novel unsupervised domain Adaptive Dense E vent Stereo (ADES), which resolves gaps between the different domains and input modalities. The proposed ADES framework adapts event-based stereo networks from abundant image datasets with ground-truth on the source domain to event datasets without ground-truth on the target domain, which is a more practical setup. Fir st, we propose a self-supervision module that trains the network on the target d omain through image reconstruction, while an artifact prediction network trained on the source domain assists in removing intermittent artifacts in the reconstr ucted image. Secondly, we utilize the feature-level normalization scheme to alig n the extracted features along the epipolar line. Finally, we present the motion -invariant consistency module to impose the consistent output between the pertur bed motion. Our experiments demonstrate that our approach achieves remarkable re sults in the adaptation ability of event-based stereo matching from the image do main.
****
Foundation Model Drives Weakly Incremental Learning for Semantic Segmentation
Chaohui Yu, Qiang Zhou, Jingliang Li, Jianlong Yuan, Zhibin Wang, Fan Wang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2023, pp. 23685-23694
Modern incremental learning for semantic segmentation methods usually learn new categories based on dense annotations. Although achieve promising results, pixel -by-pixel labeling is costly and time-consuming. Weakly incremental learning for semantic segmentation (WILSS) is a novel and attractive task, which aims at lea rning to segment new classes from cheap and widely available image-level labels. Despite the comparable results, the image-level labels can not provide details to locate each segment, which limits the performance of WILSS. This inspires us to think how to improve and effectively utilize the supervision of new classes g iven image-level labels while avoiding forgetting old ones. In this work, we pro pose a novel and data-efficient framework for WILSS, named FMWISS. Specifically, we propose pre-training based co-segmentation to distill the knowledge of compl ementary foundation models for generating dense pseudo labels. We further optimi ze the noisy pseudo masks with a teacher-student architecture, where a plug-in t eacher is optimized with a proposed dense contrastive loss. Moreover, we introdu ce memory-based copy-paste augmentation to improve the catastrophic forgetting p roblem of old classes. Extensive experiments on Pascal VOC and COCO datasets dem onstrate the superior performance of our framework, e.g., FMWISS achieves 70.7% and 73.3% in the 15-5 VOC setting, outperforming the state-of-the-art method by 3.4% and 6.1%, respectively.
****
Seeing a Rose in Five Thousand Ways
Yunzhi Zhang, Shangzhe Wu, Noah Snavely, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 962-971
What is a rose, visually? A rose comprises its intrinsics, including the distrib ution of geometry, texture, and material specific to its object category. With k nowledge of these intrinsic properties, we may render roses of different sizes a nd shapes, in different poses, and under different lighting conditions. In this work, we build a generative model that learns to capture such object intrinsics from a single image, such as a photo of a bouquet. Such an image includes multip le instances of an object type. These instances all share the same intrinsics, b ut appear different due to a combination of variance within these intrinsics and differences in extrinsic factors, such as pose and illumination. Experiments sh ow that our model successfully learns object intrinsics (distribution of geometr y, texture, and material) for a wide range of objects, each from a single Intern et image. Our method achieves superior results on multiple downstream tasks, inc luding intrinsic image decomposition, shape and image generation, view synthesis , and relighting.

****
Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos
Liao Wang, Qiang Hu, Qihan He, Ziyu Wang, Jingyi Yu, Tinne Tuytelaars, Lan Xu, Minye Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 76-87
The success of the Neural Radiance Fields (NeRFs) for modeling and free-view rendering static objects has inspired numerous attempts on dynamic scenes. Current techniques that utilize neural rendering for facilitating free-view videos (FVVs) are restricted to either offline rendering or are capable of processing only brief sequences with minimal motion. In this paper, we present a novel technique, Residual Radiance Field or ReRF, as a highly compact neural representation to achieve real-time FVV rendering on long-duration dynamic scenes. ReRF explicitly models the residual information between adjacent timestamps in the spatial-temporal feature space, with a global coordinate-based tiny MLP as the feature decoder. Specifically, ReRF employs a compact motion grid along with a residual feature grid to exploit inter-frame feature similarities. We show such a strategy can handle large motions without sacrificing quality. We further present a sequential training scheme to maintain the smoothness and the sparsity of the motion/residual grids. Based on ReRF, we design a special FVV codec that achieves three orders of magnitudes compression rate and provides a companion ReRF player to support online streaming of long-duration FVVs of dynamic scenes. Extensive experiments demonstrate the effectiveness of ReRF for compactly representing dynamic radiance fields, enabling an unprecedented free-viewpoint viewing experience in speed and quality.
****
ACSeg: Adaptive Conceptualization for Unsupervised Semantic Segmentation
Kehan Li, Zhennan Wang, Zesen Cheng, Runyi Yu, Yian Zhao, Guoli Song, Chang Liu, Li Yuan, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7162-7172
Recently, self-supervised large-scale visual pre-training models have shown great promise in representing pixel-level semantic relationships, significantly promoting the development of unsupervised dense prediction tasks, e.g., unsupervised semantic segmentation (USS). The extracted relationship among pixel-level representations typically contains rich class-aware information that semantically identical pixel embeddings in the representation space gather together to form sophisticated concepts. However, leveraging the learned models to ascertain semantically consistent pixel groups or regions in the image is non-trivial since over/under-clustering overwhelms the conceptualization procedure under various semantic distributions of different images. In this work, we investigate the pixel-level semantic aggregation in self-supervised ViT pre-trained models as image Segmentation and propose the Adaptive Conceptualization approach for USS, termed ACSeg. Concretely, we explicitly encode concepts into learnable prototypes and design the Adaptive Concept Generator (ACG), which adaptively maps these prototypes to informative concepts for each image. Meanwhile, considering the scene complexity of different images, we propose the modularity loss to optimize ACG independent of the concept number based on estimating the intensity of pixel pairs belonging to the same concept. Finally, we turn the USS task into classifying the discovered concepts in an unsupervised manner. Extensive experiments with state-of-the-art results demonstrate the effectiveness of the proposed ACSeg.
****
NeRFVS: Neural Radiance Fields for Free View Synthesis via Geometry Scaffolds
Chen Yang, Peihao Li, Zanwei Zhou, Shanxin Yuan, Bingbing Liu, Xiaokang Yang, Weichao Qiu, Wei Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16549-16558
We present NeRFVS, a novel neural radiance fields (NeRF) based method to enable free navigation in a room. NeRF achieves impressive performance in rendering images for novel views similar to the input views while suffering for novel views that are significantly different from the training views. To address this issue, we utilize the holistic priors, including pseudo depth maps and view coverage information, from neural reconstruction to guide the learning of implicit neural r

epresentations of 3D indoor scenes. Concretely, an off-the-shelf neural reconstr uction method is leveraged to generate a geometry scaffold. Then, two loss funct ions based on the holistic priors are proposed to improve the learning of NeRF: 1) A robust depth loss that can tolerate the error of the pseudo depth map to gu ide the geometry learning of NeRF; 2) A variance loss to regularize the variance of implicit neural representations to reduce the geometry and color ambiguity i n the learning procedure. These two loss functions are modulated during NeRF opt imization according to the view coverage information to reduce the negative infl uence brought by the view coverage imbalance. Extensive results demonstrate that our NeRFVS outperforms state-of-the-art view synthesis methods quantitatively a nd qualitatively on indoor scenes, achieving high-fidelity free navigation resul ts.
****
Reproducible Scaling Laws for Contrastive Language-Image Learning
Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco , Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, Jenia Jitsev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 , pp. 2818-2829
Scaling up neural networks has led to remarkable performance across a wide range of tasks. Moreover, performance often follows reliable scaling laws as a functi on of training set size, model size, and compute, which offers valuable guidance as large-scale experiments are becoming increasingly expensive. However, previo us work on scaling laws has primarily used private data & models or focused on u ni-modal language or vision learning. To address these limitations, we investiga te scaling laws for contrastive language-image pre-training (CLIP) with the publ ic LAION dataset and the open-source OpenCLIP repository. Our large-scale experi ments involve models trained on up to two billion image-text pairs and identify power law scaling for multiple downstream tasks including zero-shot classificati on, retrieval, linear probing, and end-to-end fine-tuning. We find that the trai ning distribution plays a key role in scaling laws as the OpenAI and OpenCLIP mo dels exhibit different scaling behavior despite identical model architectures an d similar training recipes. We open-source our evaluation workflow and all model s, including the largest public CLIP models, to ensure reproducibility and make scaling laws research more accessible. Source code and instructions to reproduce this study is available at https://github.com/LAION-AI/scaling-laws-openclip.
****
Similarity Metric Learning for RGB-Infrared Group Re-Identification
Jianghao Xiong, Jianhuang Lai; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2023, pp. 13662-13671
Group re-identification (G-ReID) aims to re-identify a group of people that is o bserved from non-overlapping camera systems. The existing literature has mainly addressed RGB-based problems, but RGB-infrared (RGB-IR) cross-modality matching problem has not been studied yet. In this paper, we propose a metric learning me thod Closest Permutation Matching (CPM) for RGB-IR G-ReID. We model each group a s a set of single-person features which are extracted by MPANet, then we propose the metric Closest Permutation Distance (CPD) to measure the similarity between two sets of features. CPD is invariant with order changes of group members so t hat it solves the layout change problem in G-ReID. Furthermore, we introduce the problem of G-ReID without person labels. In the weak-supervised case, we design the Relation-aware Module (RAM) that exploits visual context and relations amon g group members to produce a modality-invariant order of features in each group, with which group member features within a set can be sorted to form a robust gr oup representation against modality change. To support the study on RGB-IR G-ReI D, we construct a new large-scale RGB-IR G-ReID dataset CM-Group. The dataset co ntains 15,440 RGB images and 15,506 infrared images of 427 groups and 1,013 iden tities. Extensive experiments on the new dataset demonstrate the effectiveness o f the proposed models and the complexity of CM-Group. The code and dataset are a vailable at: https://github.com/WhollyOat/CM-Group.
****
Auto-CARD: Efficient and Robust Codec Avatar Driving for Real-Time Mobile Telepr

esence
Yonggan Fu, Yuecheng Li, Chenghui Li, Jason Saragih, Peizhao Zhang, Xiaoliang Dai, Yingyan (Celine) Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21036-21045
Real-time and robust photorealistic avatars for telepresence in AR/VR have been highly desired for enabling immersive photorealistic telepresence. However, there still exists one key bottleneck: the considerable computational expense needed to accurately infer facial expressions captured from headset-mounted cameras with a quality level that can match the realism of the avatar's human appearance. To this end, we propose a framework called Auto-CARD, which for the first time enables real-time and robust driving of Codec Avatars when exclusively using merely on-device computing resources. This is achieved by minimizing two sources of redundancy. First, we develop a dedicated neural architecture search technique called AVE-NAS for avatar encoding in AR/VR, which explicitly boosts both the searched architectures' robustness in the presence of extreme facial expressions and hardware friendliness on fast evolving AR/VR headsets. Second, we leverage the temporal redundancy in consecutively captured images during continuous rendering and develop a mechanism dubbed LATEX to skip the computation of redundant frames. Specifically, we first identify an opportunity from the linearity of the latent space derived by the avatar decoder and then propose to perform adaptive latent extrapolation for redundant frames. For evaluation, we demonstrate the efficacy of our Auto-CARD framework in real-time Codec Avatar driving settings, where we achieve a 5.05x speed-up on Meta Quest 2 while maintaining a comparable or even better animation quality than state-of-the-art avatar encoder designs.
****
Conjugate Product Graphs for Globally Optimal 2D-3D Shape Matching
Paul Roetzer, Zorah Lähner, Florian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21866-21875
We consider the problem of finding a continuous and non-rigid matching between a 2D contour and a 3D mesh. While such problems can be solved to global optimality by finding a shortest path in the product graph between both shapes, existing solutions heavily rely on unrealistic prior assumptions to avoid degenerate solutions (e.g. knowledge to which region of the 3D shape each point of the 2D contour is matched). To address this, we propose a novel 2D-3D shape matching formalism based on the conjugate product graph between the 2D contour and the 3D shape. Doing so allows us for the first time to consider higher-order costs, i.e. defined for edge chains, as opposed to costs defined for single edges. This offers substantially more flexibility, which we utilise to incorporate a local rigidity prior. By doing so, we effectively circumvent degenerate solutions and thereby obtain smoother and more realistic matchings, even when using only a one-dimensional feature descriptor. Overall, our method finds globally optimal and continuous 2D-3D matchings, has the same asymptotic complexity as previous solutions, produces state-of-the-art results for shape matching and is even capable of matching partial shapes. Our code is publicly available (https://github.com/paul0noah/sm-2D3D).
****
PromptCAL: Contrastive Affinity Learning via Auxiliary Prompts for Generalized Novel Category Discovery
Sheng Zhang, Salman Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3479-3488
Although existing semi-supervised learning models achieve remarkable success in learning with unannotated in-distribution data, they mostly fail to learn on unlabeled data sampled from novel semantic classes due to their closed-set assumption. In this work, we target a pragmatic but under-explored Generalized Novel Category Discovery (GNCD) setting. The GNCD setting aims to categorize unlabeled training data coming from known and novel classes by leveraging the information of partially labeled known classes. We propose a two-stage Contrastive Affinity Learning method with auxiliary visual Prompts, dubbed PromptCAL, to address this challenging problem. Our approach discovers reliable pairwise sample affinities t

o learn better semantic clustering of both known and novel classes for the class token and visual prompts. First, we propose a discriminative prompt regularization loss to reinforce semantic discriminativeness of prompt-adapted pre-trained vision transformer for refined affinity relationships. Besides, we propose contrastive affinity learning to calibrate semantic representations based on our iterative semi-supervised affinity graph generation method for semantically-enhanced supervision. Extensive experimental evaluation demonstrates that our PromptCAL method is more effective in discovering novel classes even with limited annotations and surpasses the current state-of-the-art on generic and fine-grained benchmarks (e.g., with nearly 11% gain on CUB-200, and 9% on ImageNet-100) on overall accuracy. Our code will be released to the public.
****
Train/Test-Time Adaptation With Retrieval
Luca Zancato, Alessandro Achille, Tian Yu Liu, Matthew Trager, Pramuditha Perera, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15911-15921
We introduce Train/Test-Time Adaptation with Retrieval (T3AR), a method to adapt models both at train and test time by means of a retrieval module and a searchable pool of external samples. Before inference, T3AR adapts a given model to the downstream task using refined pseudo-labels and a self-supervised contrastive objective function whose noise distribution leverages retrieved real samples to improve feature adaptation on the target data manifold. The retrieval of real images is key to T3AR since it does not rely solely on synthetic data augmentations to compensate for the lack of adaptation data, as typically done by other adaptation algorithms. Furthermore, thanks to the retrieval module, our method gives the user or service provider the possibility to improve model adaptation on the downstream task by incorporating further relevant data or to fully remove samples that may no longer be available due to changes in user preference after deployment. First, we show that T3AR can be used at training time to improve downstream fine-grained classification over standard fine-tuning baselines, and the fewer the adaptation data the higher the relative improvement (up to 13%). Second, we apply T3AR for test-time adaptation and show that exploiting a pool of external images at test-time leads to more robust representations over existing methods on DomainNet-126 and VISDA-C, especially when few adaptation data are available (up to 8%).
****
ProxyFormer: Proxy Alignment Assisted Point Cloud Completion With Missing Part Sensitive Transformer
Shanshan Li, Pan Gao, Xiaoyang Tan, Mingqiang Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9466-9475
Problems such as equipment defects or limited viewpoints will lead the captured point clouds to be incomplete. Therefore, recovering the complete point clouds from the partial ones plays an vital role in many practical tasks, and one of the keys lies in the prediction of the missing part. In this paper, we propose a novel point cloud completion approach namely ProxyFormer that divides point clouds into existing (input) and missing (to be predicted) parts and each part communicates information through its proxies. Specifically, we fuse information into point proxy via feature and position extractor, and generate features for missing point proxies from the features of existing point proxies. Then, in order to better perceive the position of missing points, we design a missing part sensitive transformer, which converts random normal distribution into reasonable position information, and uses proxy alignment to refine the missing proxies. It makes the predicted point proxies more sensitive to the features and positions of the missing part, and thus makes these proxies more suitable for subsequent coarse-to-fine processes. Experimental results show that our method outperforms state-of-the-art completion networks on several benchmark datasets and has the fastest inference speed.
****
Mod-Squad: Designing Mixtures of Experts As Modular Multi-Task Learners
Zitian Chen, Yikang Shen, Mingyu Ding, Zhenfang Chen, Hengshuang Zhao, Erik G. L

earned-Miller, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2023, pp. 11828-11837
Optimization in multi-task learning (MTL) is more challenging than single-task l
earning (STL), as the gradient from different tasks can be contradictory. When t
asks are related, it can be beneficial to share some parameters among them (coop
eration). However, some tasks require additional parameters with expertise in a
specific type of data or discrimination (specialization). To address the MTL cha
llenge, we propose Mod-Squad, a new model that is Modularized into groups of exp
erts (a 'Squad'). This structure allows us to formalize cooperation and speciali
zation as the process of matching experts and tasks. We optimize this matching p
rocess during the training of a single model. Specifically, we incorporate mixtu
re of experts (MoE) layers into a transformer model, with a new loss that incorp
orates the mutual dependence between tasks and experts. As a result, only a smal
l set of experts are activated for each task. This prevents the sharing of the e
ntire backbone model between all tasks, which strengthens the model, especially
when the training set size and the number of tasks scale up. More interestingly,
 for each task, we can extract the small set of experts as a standalone model th
at maintains the same performance as the large model. Extensive experiments on t
he Taskonomy dataset with 13 vision tasks and the PASCALContext dataset with 5 v
ision tasks show the superiority of our approach. The project page can be access
ed at https://vis-www.cs.umass.edu/mod-squad.
****
Learning Customized Visual Models With Retrieval-Augmented Knowledge
Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, Chunyu
an Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2023, pp. 15148-15158
Image-text contrastive learning models such as CLIP have demonstrated strong tas
k transfer ability. The high generality and usability of these visual models is
achieved via a web-scale data collection process to ensure broad concept coverag
e, followed by expensive pre-training to feed all the knowledge into model weigh
ts. Alternatively, we propose REACT, REtrieval-Augmented CusTomization, a framew
ork to acquire the relevant web knowledge to build customized visual models for
target domains. We retrieve the most relevant image-text pairs (3% of CLIP pre-
training data) from the web-scale database as external knowledge and propose to
customize the model by only training new modularized blocks while freezing all t
he original weights. The effectiveness of REACT is demonstrated via extensive ex
periments on classification, retrieval, detection and segmentation tasks, includ
ing zero, few, and full-shot settings. Particularly, on the zero-shot classifica
tion task, compared with CLIP, it achieves up to 5.4% improvement on ImageNet an
d 3.7% on the ELEVATER benchmark (20 datasets).
****
Multi-Realism Image Compression With a Conditional Generator
Eirikur Agustsson, David Minnen, George Toderici, Fabian Mentzer; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 22324-22333
By optimizing the rate-distortion-realism trade-off, generative compression appr
oaches produce detailed, realistic images, even at low bit rates, instead of the
 blurry reconstructions produced by rate-distortion optimized models. However, p
revious methods do not explicitly control how much detail is synthesized, which
results in a common criticism of these methods: users might be worried that a mi
sleading reconstruction far from the input image is generated. In this work, we
alleviate these concerns by training a decoder that can bridge the two regimes a
nd navigate the distortion-realism trade-off. From a single compressed represent
ation, the receiver can decide to either reconstruct a low mean squared error re
construction that is close to the input, a realistic reconstruction with high pe
rceptual quality, or anything in between. With our method, we set a new state-of
-the-art in distortion-realism, pushing the frontier of achievable distortion-re
alism pairs, i.e., our method achieves better distortions at high realism and be
tter realism at low distortion than ever before.
****

Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks
Jierun Chen, Shiu-hong Kao, Hao He, Weipeng Zhuo, Song Wen, Chul-Ho Lee, S.-H. Gary Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12021-12031
To design fast neural networks, many works have been focusing on reducing the number of floating-point operations (FLOPs). We observe that such reduction in FLOPs, however, does not necessarily lead to a similar level of reduction in latency. This mainly stems from inefficiently low floating-point operations per second (FLOPS). To achieve faster networks, we revisit popular operators and demonstrate that such low FLOPS is mainly due to frequent memory access of the operators, especially the depthwise convolution. We hence propose a novel partial convolution (PConv) that extracts spatial features more efficiently, by cutting down redundant computation and memory access simultaneously. Building upon our PConv, we further propose FasterNet, a new family of neural networks, which attains substantially higher running speed than others on a wide range of devices, without compromising on accuracy for various vision tasks. For example, on ImageNet-1k, our tiny FasterNet-T0 is 2.8x, 3.3x, and 2.4x faster than MobileViT-XXS on GPU, CPU, and ARM processors, respectively, while being 2.9% more accurate. Our large FasterNet-L achieves impressive 83.5% top-1 accuracy, on par with the emerging Swin-B, while having 36% higher inference throughput on GPU, as well as saving 37% compute time on CPU. Code is available at https://github.com/JierunChen/FasterNet.
****
A Unified Spatial-Angular Structured Light for Single-View Acquisition of Shape and Reflectance
Xianmin Xu, Yuxin Lin, Haoyang Zhou, Chong Zeng, Yaxin Yu, Kun Zhou, Hongzhi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 206-215
We propose a unified structured light, consisting of an LED array and an LCD mask, for high-quality acquisition of both shape and reflectance from a single view. For geometry, one LED projects a set of learned mask patterns to accurately encode spatial information; the decoded results from multiple LEDs are then aggregated to produce a final depth map. For appearance, learned light patterns are cast through a transparent mask to efficiently probe angularly-varying reflectance. Per-point BRDF parameters are differentiably optimized with respect to corresponding measurements, and stored in texture maps as the final reflectance. We establish a differentiable pipeline for the joint capture to automatically optimize both the mask and light patterns towards optimal acquisition quality. The effectiveness of our light is demonstrated with a wide variety of physical objects. Our results compare favorably with state-of-the-art techniques.
****
Best of Both Worlds: Multimodal Contrastive Learning With Tabular and Imaging Data
Paul Hager, Martin J. Menten, Daniel Rueckert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23924-23935
Medical datasets and especially biobanks, often contain extensive tabular data with rich clinical information in addition to images. In practice, clinicians typically have less data, both in terms of diversity and scale, but still wish to deploy deep learning solutions. Combined with increasing medical dataset sizes and expensive annotation costs, the necessity for unsupervised methods that can pretrain multimodally and predict unimodally has risen. To address these needs, we propose the first self-supervised contrastive learning framework that takes advantage of images and tabular data to train unimodal encoders. Our solution combines SimCLR and SCARF, two leading contrastive learning strategies, and is simple and effective. In our experiments, we demonstrate the strength of our framework by predicting risks of myocardial infarction and coronary artery disease (CAD) using cardiac MR images and 120 clinical features from 40,000 UK Biobank subjects. Furthermore, we show the generalizability of our approach to natural images using the DVM car advertisement dataset. We take advantage of the high interpretability of tabular data and through attribution and ablation experiments find tha

t morphometric tabular features, describing size and shape, have outsized importance during the contrastive learning process and improve the quality of the learned embeddings. Finally, we introduce a novel form of supervised contrastive learning, label as a feature (LaaF), by appending the ground truth label as a tabular feature during multimodal pretraining, outperforming all supervised contrastive baselines.
****

On the Difficulty of Unpaired Infrared-to-Visible Video Translation: Fine-Grained Content-Rich Patches Transfer

Zhenjie Yu, Shuang Li, Yirui Shen, Chi Harold Liu, Shuigen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1631-1640

Explicit visible videos can provide sufficient visual information and facilitate vision applications. Unfortunately, the image sensors of visible cameras are sensitive to light conditions like darkness or overexposure. To make up for this, recently, infrared sensors capable of stable imaging have received increasing attention in autonomous driving and monitoring. However, most prosperous vision models are still trained on massive clear visible data, facing huge visual gaps when deploying to infrared imaging scenarios. In such cases, transferring the infrared video to a distinct visible one with fine-grained semantic patterns is a worthwhile endeavor. Previous works improve the outputs by equally optimizing each patch on the translated visible results, which is unfair for enhancing the details on content-rich patches due to the long-tail effect of pixel distribution. Here we propose a novel CPTrans framework to tackle the challenge via balancing gradients of different patches, achieving the fine-grained Content-rich Patches Transferring. Specifically, the content-aware optimization module encourages model optimization along gradients of target patches, ensuring the improvement of visual details. Additionally, the content-aware temporal normalization module enforces the generator to be robust to the motions of target patches. Moreover, we extend the existing dataset InfraredCity to more challenging adverse weather conditions (rain and snow), dubbed as InfraredCity-Adverse. Extensive experiments show that the proposed CPTrans achieves state-of-the-art performance under diverse scenes while requiring less training time than competitive methods.
****

Masked Images Are Counterfactual Samples for Robust Fine-Tuning

Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, Liang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20301-20310

Deep learning models are challenged by the distribution shift between the training data and test data. Recently, the large models pre-trained on diverse data have demonstrated unprecedented robustness to various distribution shifts. However, fine-tuning these models can lead to a trade-off between in-distribution (ID) performance and out-of-distribution (OOD) robustness. Existing methods for tackling this trade-off do not explicitly address the OOD robustness problem. In this paper, based on causal analysis of the aforementioned problems, we propose a novel fine-tuning method, which uses masked images as counterfactual samples that help improve the robustness of the fine-tuning model. Specifically, we mask either the semantics-related or semantics-unrelated patches of the images based on class activation map to break the spurious correlation, and refill the masked patches with patches from other images. The resulting counterfactual samples are used in feature-based distillation with the pre-trained model. Extensive experiments verify that regularizing the fine-tuning with the proposed masked images can achieve a better trade-off between ID and OOD performance, surpassing previous methods on the OOD performance. Our code is available at https://github.com/Coxy7/robust-finetuning.
****

StepFormer: Self-Supervised Step Discovery and Localization in Instructional Videos

Nikita Dvornik, Isma Hadji, Ran Zhang, Konstantinos G. Derpanis, Richard P. Wildes, Allan D. Jepson; Proceedings of the IEEE/CVF Conference on Computer Vision a

nd Pattern Recognition (CVPR), 2023, pp. 18952-18961
Instructional videos are an important resource to learn procedural tasks from human demonstrations. However, the instruction steps in such videos are typically short and sparse, with most of the video being irrelevant to the procedure. This motivates the need to temporally localize the instruction steps in such videos, i.e. the task called key-step localization. Traditional methods for key-step localization require video-level human annotations and thus do not scale to large datasets. In this work, we tackle the problem with no human supervision and introduce StepFormer, a self-supervised model that discovers and localizes instruction steps in a video. StepFormer is a transformer decoder that attends to the video with learnable queries, and produces a sequence of slots capturing the key-steps in the video. We train our system on a large dataset of instructional videos, using their automatically-generated subtitles as the only source of supervision. In particular, we supervise our system with a sequence of text narrations using an order-aware loss function that filters out irrelevant phrases. We show that our model outperforms all previous unsupervised and weakly-supervised approaches on step detection and localization by a large margin on three challenging benchmarks. Moreover, our model demonstrates an emergent property to solve zero-shot multi-step localization and outperforms all relevant baselines at this task.
****
Learning Procedure-Aware Video Representation From Instructional Videos and Their Narrations
Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, Yin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14825-14835
The abundance of instructional videos and their narrations over the Internet offers an exciting avenue for understanding procedural activities. In this work, we propose to learn video representation that encodes both action steps and their temporal ordering, based on a large-scale dataset of web instructional videos and their narrations, without using human annotations. Our method jointly learns a video representation to encode individual step concepts, and a deep probabilistic model to capture both temporal dependencies and immense individual variations in the step ordering. We empirically demonstrate that learning temporal ordering not only enables new capabilities for procedure reasoning, but also reinforces the recognition of individual steps. Our model significantly advances the state-of-the-art results on step classification (+2.8% / +3.3% on COIN / EPIC-Kitchens) and step forecasting (+7.4% on COIN). Moreover, our model attains promising results in zero-shot inference for step classification and forecasting, as well as in predicting diverse and plausible steps for incomplete procedures. Our code is available at https://github.com/facebookresearch/ProcedureVRL.
****
Open Vocabulary Semantic Segmentation With Patch Aligned Contrastive Learning
Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H.S. Torr, Ser-Nam Lim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19413-19423
We introduce Patch Aligned Contrastive Learning (PACL), a modified compatibility function for CLIP's contrastive loss, intending to train an alignment between the patch tokens of the vision encoder and the CLS token of the text encoder. With such an alignment, a model can identify regions of an image corresponding to a given text input, and therefore transfer seamlessly to the task of open vocabulary semantic segmentation without requiring any segmentation annotations during training. Using pre-trained CLIP encoders with PACL, we are able to set the state-of-the-art on the task of open vocabulary zero-shot segmentation on 4 different segmentation benchmarks: Pascal VOC, Pascal Context, COCO Stuff and ADE20K. Furthermore, we show that PACL is also applicable to image-level predictions and when used with a CLIP backbone, provides a general improvement in zero-shot classification accuracy compared to CLIP, across a suite of 12 image classification datasets.
****
CLIP the Gap: A Single Domain Generalization Approach for Object Detection

Vidit Vidit, Martin Engilberge, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3219-3229
Single Domain Generalization (SDG) tackles the problem of training a model on a single source domain so that it generalizes to any unseen target domain. While this has been well studied for image classification, the literature on SDG object detection remains almost non-existent. To address the challenges of simultaneously learning robust object localization and representation, we propose to leverage a pre-trained vision-language model to introduce semantic domain concepts via textual prompts. We achieve this via a semantic augmentation strategy acting on the features extracted by the detector backbone, as well as a text-based classification loss. Our experiments evidence the benefits of our approach, outperforming by 10% the only existing SDG object detection method, Single-DGOD[49], on their own diverse weather-driving benchmark.
****

Co-Training 2L Submodels for Visual Recognition
Hugo Touvron, Matthieu Cord, Maxime Oquab, Piotr Bojanowski, Jakob Verbeek, Hervé Jégou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11701-11710
This paper introduces submodel co-training, a regularization method related to co-training, self-distillation and stochastic depth. Given a neural network to be trained, for each sample we implicitly instantiate two altered networks, "submodels", with stochastic depth: i.e. activating only a subset of the layers and skipping others. Each network serves as a soft teacher to the other, by providing a cross-entropy loss that complements the regular softmax cross-entropy loss provided by the one-hot label. Our approach, dubbed "cosub", uses a single set of weights, and does not involve a pre-trained external model or temporal averaging. Experimentally, we show that submodel co-training is effective to train backbones for recognition tasks such as image classification and semantic segmentation, and that our approach is compatible with multiple recent architectures, including RegNet, PiT, and Swin. We report new state-of-the-art results for vision transformers trained on ImageNet only. For instance, a ViT-B pre-trained with cosub on Imagenet-21k achieves 87.4% top-1 acc. on Imagenet-val.
****

On the Importance of Accurate Geometry Data for Dense 3D Vision Tasks
HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Aleš Leonardis, Nassir Navab, Benjamin Busam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 780-791
Learning-based methods to solve dense 3D vision problems typically train on 3D sensor data. The respectively used principle of measuring distances provides advantages and drawbacks. These are typically not compared nor discussed in the literature due to a lack of multi-modal datasets. Texture-less regions are problematic for structure from motion and stereo, reflective material poses issues for active sensing, and distances for translucent objects are intricate to measure with existing hardware. Training on inaccurate or corrupt data induces model bias and hampers generalisation capabilities. These effects remain unnoticed if the sensor measurement is considered as ground truth during the evaluation. This paper investigates the effect of sensor errors for the dense 3D vision tasks of depth estimation and reconstruction. We rigorously show the significant impact of sensor characteristics on the learned predictions and notice generalisation issues arising from various technologies in everyday household environments. For evaluation, we introduce a carefully designed dataset comprising measurements from commodity sensors, namely D-ToF, I-ToF, passive/active stereo, and monocular RGB+P. Our study quantifies the considerable sensor noise impact and paves the way to improved dense vision estimates and targeted data fusion.
****

Camouflaged Instance Segmentation via Explicit De-Camouflaging
Naisong Luo, Yuwen Pan, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17918-17927

Camouflaged Instance Segmentation (CIS) aims at predicting the instance-level masks of camouflaged objects, which are usually the animals in the wild adapting their appearance to match the surroundings. Previous instance segmentation methods perform poorly on this task as they are easily disturbed by the deceptive camouflage. To address these challenges, we propose a novel De-camouflaging Network (DCNet) including a pixel-level camouflage decoupling module and an instance-level camouflage suppression module. The proposed DCNet enjoys several merits. First, the pixel-level camouflage decoupling module can extract camouflage characteristics based on the Fourier transformation. Then a difference attention mechanism is proposed to eliminate the camouflage characteristics while reserving target object characteristics in the pixel feature. Second, the instance-level camouflage suppression module can aggregate rich instance information from pixels by use of instance prototypes. To mitigate the effect of background noise during segmentation, we introduce some reliable reference points to build a more robust similarity measurement. With the aid of these two modules, our DCNet can effectively model de-camouflaging and achieve accurate segmentation for camouflaged instances. Extensive experimental results on two benchmarks demonstrate that our DCNet performs favorably against state-of-the-art CIS methods, e.g., with more than 5% performance gains on COD10K and NC4K datasets in average precision.
****
Understanding Masked Autoencoders via Hierarchical Latent Variable Models
Lingjing Kong, Martin Q. Ma, Guangyi Chen, Eric P. Xing, Yuejie Chi, Louis-Philippe Morency, Kun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7918-7928
Masked autoencoder (MAE), a simple and effective self-supervised learning framework based on the reconstruction of masked image regions, has recently achieved prominent success in a variety of vision tasks. Despite the emergence of intriguing empirical observations on MAE, a theoretically principled understanding is still lacking. In this work, we formally characterize and justify existing empirical insights and provide theoretical guarantees of MAE. We formulate the underlying data-generating process as a hierarchical latent variable model, and show that under reasonable assumptions, MAE provably identifies a set of latent variables in the hierarchical model, explaining why MAE can extract high-level information from pixels. Further, we show how key hyperparameters in MAE (the masking ratio and the patch size) determine which true latent variables to be recovered, therefore influencing the level of semantic information in the representation. Specifically, extremely large or small masking ratios inevitably lead to low-level representations. Our theory offers coherent explanations of existing empirical observations and provides insights for potential empirical improvements and fundamental limitations of the masked-reconstruction paradigm. We conduct extensive experiments to validate our theoretical insights.
****
K-Planes: Explicit Radiance Fields in Space, Time, and Appearance
Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12479-12488
We introduce k-planes, a white-box model for radiance fields in arbitrary dimensions. Our model uses d-choose-2 planes to represent a d-dimensional scene, providing a seamless way to go from static (d=3) to dynamic (d=4) scenes. This planar factorization makes adding dimension-specific priors easy, e.g. temporal smoothness and multi-resolution spatial structure, and induces a natural decomposition of static and dynamic components of a scene. We use a linear feature decoder with a learned color basis that yields similar performance as a nonlinear black-box MLP decoder. Across a range of synthetic and real, static and dynamic, fixed and varying appearance scenes, k-planes yields competitive and often state-of-the-art reconstruction fidelity with low memory usage, achieving 1000x compression over a full 4D grid, and fast optimization with a pure PyTorch implementation. For video results and code, please see sarafridov.github.io/K-Planes.
****
Multi-Mode Online Knowledge Distillation for Self-Supervised Visual Representati

on Learning
Kaiyou Song, Jin Xie, Shan Zhang, Zimeng Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11848-11857
Self-supervised learning (SSL) has made remarkable progress in visual representation learning. Some studies combine SSL with knowledge distillation (SSL-KD) to boost the representation learning performance of small models. In this study, we propose a Multi-mode Online Knowledge Distillation method (MOKD) to boost self-supervised visual representation learning. Different from existing SSL-KD methods that transfer knowledge from a static pre-trained teacher to a student, in MOKD, two different models learn collaboratively in a self-supervised manner. Specifically, MOKD consists of two distillation modes: self-distillation and cross-distillation modes. Among them, self-distillation performs self-supervised learning for each model independently, while cross-distillation realizes knowledge interaction between different models. In cross-distillation, a cross-attention feature search strategy is proposed to enhance the semantic feature alignment between different models. As a result, the two models can absorb knowledge from each other to boost their representation learning performance. Extensive experimental results on different backbones and datasets demonstrate that two heterogeneous models can benefit from MOKD and outperform their independently trained baseline. In addition, MOKD also outperforms existing SSL-KD methods for both the student and teacher models.
****
Unbalanced Optimal Transport: A Unified Framework for Object Detection
Henri De Plaen, Pierre-François De Plaen, Johan A. K. Suykens, Marc Proesmans, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3198-3207
During training, supervised object detection tries to correctly match the predicted bounding boxes and associated classification scores to the ground truth. This is essential to determine which predictions are to be pushed towards which solutions, or to be discarded. Popular matching strategies include matching to the closest ground truth box (mostly used in combination with anchors), or matching via the Hungarian algorithm (mostly used in anchor-free methods). Each of these strategies comes with its own properties, underlying losses, and heuristics. We show how Unbalanced Optimal Transport unifies these different approaches and opens a whole continuum of methods in between. This allows for a finer selection of the desired properties. Experimentally, we show that training an object detection model with Unbalanced Optimal Transport is able to reach the state-of-the-art both in terms of Average Precision and Average Recall as well as to provide a faster initial convergence. The approach is well suited for GPU implementation, which proves to be an advantage for large-scale models.
****
Viewpoint Equivariance for Multi-View 3D Object Detection
Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9213-9222
3D object detection from visual sensors is a cornerstone capability of robotic systems. State-of-the-art methods focus on reasoning and decoding object bounding boxes from multi-view camera input. In this work we gain intuition from the integral role of multi-view consistency in 3D scene understanding and geometric learning. To this end, we introduce VEDet, a novel 3D object detection framework that exploits 3D multi-view geometry to improve localization through viewpoint awareness and equivariance. VEDet leverages a query-based transformer architecture and encodes the 3D scene by augmenting image features with positional encodings from their 3D perspective geometry. We design view-conditioned queries at the output level, which enables the generation of multiple virtual frames during training to learn viewpoint equivariance by enforcing multi-view consistency. The multi-view geometry injected at the input level as positional encodings and regularized at the loss level provides rich geometric cues for 3D object detection, leading to state-of-the-art performance on the nuScenes benchmark. The code and model are made available at https://github.com/TRI-ML/VEDet.

****
Photo Pre-Training, but for Sketch
Ke Li, Kaiyue Pang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2754-2764
The sketch community has faced up to its unique challenges over the years, that of data scarcity however still remains the most significant to date. This lack of sketch data has imposed on the community a few "peculiar" design choices -- the most representative of them all is perhaps the coerced utilisation of photo-based pre-training (i.e., no sketch), for many core tasks that otherwise dictates specific sketch understanding. In this paper, we ask just the one question -- can we make such photo-based pre-training, to actually benefit sketch? Our answer lies in cultivating the topology of photo data learned at pre-training, and use that as a "free" source of supervision for downstream sketch tasks. In particular, we use fine-grained sketch-based image retrieval (FG-SBIR), one of the most studied and data-hungry sketch tasks, to showcase our new perspective on pre-training. In this context, the topology-informed supervision learned from photos act as a constraint that take effect at every fine-tuning step -- neighbouring photos in the pre-trained model remain neighbours under each FG-SBIR updates. We further portray this neighbourhood consistency constraint as a photo ranking problem and formulate it into a neat cross-modal triplet loss. We also show how this target is better leveraged as a meta objective rather than optimised in parallel with the main FG-SBIR objective. With just this change on pre-training, we beat all previously published results on all five product-level FG-SBIR benchmarks with significant margins (sometimes >10%). And the most beautiful thing, as we note, is such gigantic leap is made possible with just a few extra lines of code! Our implementation is available at https://github.com/KeLi-SketchX/Photo-Pre-Training-But-for-Sketch
****
NeuralPCI: Spatio-Temporal Neural Field for 3D Point Cloud Multi-Frame Non-Linear Interpolation
Zehan Zheng, Danni Wu, Ruisi Lu, Fan Lu, Guang Chen, Changjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 909-918
In recent years, there has been a significant increase in focus on the interpolation task of computer vision. Despite the tremendous advancement of video interpolation, point cloud interpolation remains insufficiently explored. Meanwhile, the existence of numerous nonlinear large motions in real-world scenarios makes the point cloud interpolation task more challenging. In light of these issues, we present NeuralPCI: an end-to-end 4D spatio-temporal Neural field for 3D Point Cloud Interpolation, which implicitly integrates multi-frame information to handle nonlinear large motions for both indoor and outdoor scenarios. Furthermore, we construct a new multi-frame point cloud interpolation dataset called NL-Drive for large nonlinear motions in autonomous driving scenes to better demonstrate the superiority of our method. Ultimately, NeuralPCI achieves state-of-the-art performance on both DHB (Dynamic Human Bodies) and NL-Drive datasets. Beyond the interpolation task, our method can be naturally extended to point cloud extrapolation, morphing, and auto-labeling, which indicates substantial potential in other domains. Codes are available at https://github.com/ispc-lab/NeuralPCI.
****
Bidirectional Cross-Modal Knowledge Exploration for Video Recognition With Pre-Trained Vision-Language Models
Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6620-6630
Vision-language models (VLMs) pre-trained on large-scale image-text pairs have demonstrated impressive transferability on various visual tasks. Transferring knowledge from such powerful VLMs is a promising direction for building effective video recognition models. However, current exploration in this field is still limited. We believe that the greatest value of pre-trained VLMs lies in building a bridge between visual and textual domains. In this paper, we propose a novel fra

mework called BIKE, which utilizes the cross-modal bridge to explore bidirection al knowledge: i) We introduce the Video Attribute Association mechanism, which l everages the Video-to-Text knowledge to generate textual auxiliary attributes fo r complementing video recognition. ii) We also present a Temporal Concept Spotti ng mechanism that uses the Text-to-Video expertise to capture temporal saliency in a parameter-free manner, leading to enhanced video representation. Extensive studies on six popular video datasets, including Kinetics-400 & 600, UCF-101, HM DB-51, ActivityNet and Charades, show that our method achieves state-of-the-art performance in various recognition scenarios, such as general, zero-shot, and fe w-shot video recognition. Our best model achieves a state-of-the-art accuracy of 88.6% on the challenging Kinetics-400 using the released CLIP model. The code i s available at https://github.com/whwu95/BIKE.
****
Adaptive Plasticity Improvement for Continual Learning
Yan-Shuo Liang, Wu-Jun Li; Proceedings of the IEEE/CVF Conference on Computer Vi sion and Pattern Recognition (CVPR), 2023, pp. 7816-7825
Many works have tried to solve the catastrophic forgetting (CF) problem in conti nual learning (lifelong learning). However, pursuing non-forgetting on old tasks may damage the model's plasticity for new tasks. Although some methods have bee n proposed to achieve stability-plasticity trade-off, no methods have considered evaluating a model's plasticity and improving plasticity adaptively for a new t ask. In this work, we propose a new method, called adaptive plasticity improveme nt (API), for continual learning. Besides the ability to overcome CF on old task s, API also tries to evaluate the model's plasticity and then adaptively improve the model's plasticity for learning a new task if necessary. Experiments on sev eral real datasets show that API can outperform other state-of-the-art baselines in terms of both accuracy and memory usage.
****
Pic2Word: Mapping Pictures to Words for Zero-Shot Composed Image Retrieval
Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko , Tomas Pfister; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2023, pp. 19305-19314
In Composed Image Retrieval (CIR), a user combines a query image with text to de scribe their intended target. Existing methods rely on supervised learning of CI R models using labeled triplets consisting of the query image, text specificatio n, and the target image. Labeling such triplets is expensive and hinders broad a pplicability of CIR. In this work, we propose to study an important task, Zero-S hot Composed Image Retrieval (ZS-CIR), whose goal is to build a CIR model withou t requiring labeled triplets for training. To this end, we propose a novel metho d, called Pic2Word, that requires only weakly labeled image-caption pairs and un labeled image datasets to train. Unlike existing supervised CIR models, our mode l trained on weakly labeled or unlabeled datasets shows strong generalization ac ross diverse ZS-CIR tasks, e.g., attribute editing, object composition, and doma in conversion. Our approach outperforms several supervised CIR methods on the co mmon CIR benchmark, CIRR and Fashion-IQ.
****
MMANet: Margin-Aware Distillation and Modality-Aware Regularization for Incomple te Multimodal Learning
Shicai Wei, Chunbo Luo, Yang Luo; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2023, pp. 20039-20049
Multimodal learning has shown great potentials in numerous scenes and attracts i ncreasing interest recently. However, it often encounters the problem of missing modality data and thus suffers severe performance degradation in practice. To t his end, we propose a general framework called MMANet to assist incomplete multi modal learning. It consists of three components: the deployment network used for inference, the teacher network transferring comprehensive multimodal informatio n to the deployment network, and the regularization network guiding the deployme nt network to balance weak modality combinations. Specifically, we propose a nov el margin-aware distillation (MAD) to assist the information transfer by weighin g the sample contribution with the classification uncertainty. This encourages t

he deployment network to focus on the samples near decision boundaries and acquire the refined inter-class margin. Besides, we design a modality-aware regularization (MAR) algorithm to mine the weak modality combinations and guide the regularization network to calculate prediction loss for them. This forces the deployment network to improve its representation ability for the weak modality combinations adaptively. Finally, extensive experiments on multimodal classification and segmentation tasks demonstrate that our MMANet outperforms the state-of-the-art significantly.
****
Putting People in Their Place: Affordance-Aware Human Insertion Into Scenes
Sumith Kulal, Tim Brooks, Alex Aiken, Jiajun Wu, Jimei Yang, Jingwan Lu, Alexei A. Efros, Krishna Kumar Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17089-17099
We study the problem of inferring scene affordances by presenting a method for realistically inserting people into scenes. Given a scene image with a marked region and an image of a person, we insert the person into the scene while respecting the scene affordances. Our model can infer the set of realistic poses given the scene context, re-pose the reference person, and harmonize the composition. We set up the task in a self-supervised fashion by learning to re- pose humans in video clips. We train a large-scale diffusion model on a dataset of 2.4M video clips that produces diverse plausible poses while respecting the scene context. Given the learned human-scene composition, our model can also hallucinate realistic people and scenes when prompted without conditioning and also enables interactive editing. We conduct quantitative evaluation and show that our method synthesizes more realistic human appearance and more natural human-scene interactions when compared to prior work.
****
3D Neural Field Generation Using Triplane Diffusion
J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20875-20886
Diffusion models have emerged as the state-of-the-art for image generation, among other tasks. Here, we present an efficient diffusion-based model for 3D-aware generation of neural fields. Our approach pre-processes training data, such as ShapeNet meshes, by converting them to continuous occupancy fields and factoring them into a set of axis-aligned triplane feature representations. Thus, our 3D training scenes are all represented by 2D feature planes, and we can directly train existing 2D diffusion models on these representations to generate 3D neural fields with high quality and diversity, outperforming alternative approaches to 3D-aware generation. Our approach requires essential modifications to existing triplane factorization pipelines to make the resulting features easy to learn for the diffusion model. We demonstrate state-of-the-art results on 3D generation on several object classes from ShapeNet.
****
Regularized Vector Quantization for Tokenized Image Synthesis
Jiahui Zhang, Fangneng Zhan, Christian Theobalt, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18467-18476
Quantizing images into discrete representations has been a fundamental problem in unified generative modeling. Predominant approaches learn the discrete representation either in a deterministic manner by selecting the best-matching token or in a stochastic manner by sampling from a predicted distribution. However, deterministic quantization suffers from severe codebook collapse and misaligned inference stage while stochastic quantization suffers from low codebook utilization and perturbed reconstruction objective. This paper presents a regularized vector quantization framework that allows to mitigate above issues effectively by applying regularization from two perspectives. The first is a prior distribution regularization which measures the discrepancy between a prior token distribution and predicted token distribution to avoid codebook collapse and low codebook utilization. The second is a stochastic mask regularization that introduces stochasti

city during quantization to strike a good balance between inference stage misalignment and unperturbed reconstruction objective. In addition, we design a probabilistic contrastive loss which serves as a calibrated metric to further mitigate the perturbed reconstruction objective. Extensive experiments show that the proposed quantization framework outperforms prevailing vector quantizers consistently across different generative models including auto-regressive models and diffusion models.
****
Semantic Scene Completion With Cleaner Self
Fengyun Wang, Dong Zhang, Hanwang Zhang, Jinhui Tang, Qianru Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 867-877
Semantic Scene Completion (SSC) transforms an image of single-view depth and/or RGB 2D pixels into 3D voxels, each of whose semantic labels are predicted. SSC is a well-known ill-posed problem as the prediction model has to "imagine" what is behind the visible surface, which is usually represented by Truncated Signed Distance Function (TSDF). Due to the sensory imperfection of the depth camera, most existing methods based on the noisy TSDF estimated from depth values suffer from 1) incomplete volumetric predictions and 2) confused semantic labels. To this end, we use the ground-truth 3D voxels to generate a perfect visible surface, called TSDF-CAD, and then train a "cleaner" SSC model. As the model is noise-free, it is expected to focus more on the "imagination" of unseen voxels. Then, we propose to distill the intermediate "cleaner" knowledge into another model with noisy TSDF input. In particular, we use the 3D occupancy feature and the semantic relations of the "cleaner self" to supervise the counterparts of the "noisy self" to respectively address the above two incorrect predictions. Experimental results validate that the proposed method improves the noisy counterparts with 3.1% IoU and 2.2% mIoU for measuring scene completion and SSC, and also achieves new state-of-the-art accuracy on the popular NYU dataset. The code is available at https://github.com/fereenwong/CleanerS.
****
Improving Image Recognition by Retrieving From Web-Scale Image-Text Data
Ahmet Iscen, Alireza Fathi, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19295-19304
Retrieval augmented models are becoming increasingly popular for computer vision tasks after their recent success in NLP problems. The goal is to enhance the recognition capabilities of the model by retrieving similar examples for the visual input from an external memory set. In this work, we introduce an attention-based memory module, which learns the importance of each retrieved example from the memory. Compared to existing approaches, our method removes the influence of the irrelevant retrieved examples, and retains those that are beneficial to the input query. We also thoroughly study various ways of constructing the memory dataset. Our experiments show the benefit of using a massive-scale memory dataset of 1B image-text pairs, and demonstrate the performance of different memory representations. We evaluate our method in three different classification tasks, namely long-tailed recognition, learning with noisy labels, and fine-grained classification, and show that it achieves state-of-the-art accuracies in ImageNet-LT, Places-LT and Webvision datasets.
****
Deep Factorized Metric Learning
Chengkun Wang, Wenzhao Zheng, Junlong Li, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7672-7682
Learning a generalizable and comprehensive similarity metric to depict the semantic discrepancies between images is the foundation of many computer vision tasks. While existing methods approach this goal by learning an ensemble of embeddings with diverse objectives, the backbone network still receives a mix of all the training signals. Differently, we propose a deep factorized metric learning method (DFML) to factorize the training signal and employ different samples to train various components of the backbone network. We factorize the network to differe

nt sub-blocks and devise a learnable router to adaptively allocate the training samples to each sub-block with the objective to capture the most information. The metric model trained by DFML captures different characteristics with different sub-blocks and constitutes a generalizable metric when using all the sub-blocks. The proposed DFML achieves state-of-the-art performance on all three benchmarks for deep metric learning including CUB-200-2011, Cars196, and Stanford Online Products. We also generalize DFML to the image classification task on ImageNet-1K and observe consistent improvement in accuracy/computation trade-off. Specifically, we improve the performance of ViT-B on ImageNet (+0.2% accuracy) with less computation load (-24% FLOPs).
****

High-Fidelity 3D Face Generation From Natural Language Descriptions
Menghua Wu, Hao Zhu, Linjia Huang, Yiyu Zhuang, Yuanxun Lu, Xun Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4521-4530
Synthesizing high-quality 3D face models from natural language descriptions is very valuable for many applications, including avatar creation, virtual reality, and telepresence. However, little research ever tapped into this task. We argue the major obstacle lies in 1) the lack of high-quality 3D face data with descriptive text annotation, and 2) the complex mapping relationship between descriptive language space and shape/appearance space. To solve these problems, we build DESCRIBE3D dataset, the first large-scale dataset with fine-grained text descriptions for text-to-3D face generation task. Then we propose a two-stage framework to first generate a 3D face that matches the concrete descriptions, then optimize the parameters in the 3D shape and texture space with abstract description to refine the 3D face model. Extensive experimental results show that our method can produce a faithful 3D face that conforms to the input descriptions with higher accuracy and quality than previous methods. The code and DESCRIBE3D dataset are released at https://github.com/zhuhao-nju/describe3d.
****

A Generalized Framework for Video Instance Segmentation
Miran Heo, Sukjun Hwang, Jeongseok Hyun, Hanjung Kim, Seoung Wug Oh, Joon-Young Lee, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14623-14632
The handling of long videos with complex and occluded sequences has recently emerged as a new challenge in the video instance segmentation (VIS) community. However, existing methods have limitations in addressing this challenge. We argue that the biggest bottleneck in current approaches is the discrepancy between training and inference. To effectively bridge this gap, we propose a Generalized framework for VIS, namely GenVIS, that achieves state-of-the-art performance on challenging benchmarks without designing complicated architectures or requiring extra post-processing. The key contribution of GenVIS is the learning strategy, which includes a query-based training pipeline for sequential learning with a novel target label assignment. Additionally, we introduce a memory that effectively acquires information from previous states. Thanks to the new perspective, which focuses on building relationships between separate frames or clips, GenVIS can be flexibly executed in both online and semi-online manner. We evaluate our approach on popular VIS benchmarks, achieving state-of-the-art results on YouTube-VIS 2019/2021/2022 and Occluded VIS (OVIS). Notably, we greatly outperform the state-of-the-art on the long VIS benchmark (OVIS), improving 5.6 AP with ResNet-50 backbone. Code is available at https://github.com/miranheo/GenVIS.
****

Multi-Level Logit Distillation
Ying Jin, Jiaqi Wang, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24276-24285
Knowledge Distillation (KD) aims at distilling the knowledge from the large teacher model to a lightweight student model. Mainstream KD methods can be divided into two categories, logit distillation, and feature distillation. The former is easy to implement, but inferior in performance, while the latter is not applicable to some practical circumstances due to concerns such as privacy and safety. T

owards this dilemma, in this paper, we explore a stronger logit distillation method via making better utilization of logit outputs. Concretely, we propose a simple yet effective approach to logit distillation via multi-level prediction alignment. Through this framework, the prediction alignment is not only conducted at the instance level, but also at the batch and class level, through which the student model learns instance prediction, input correlation, and category correlation simultaneously. In addition, a prediction augmentation mechanism based on model calibration further boosts the performance. Extensive experiment results validate that our method enjoys consistently higher performance than previous logit distillation methods, and even reaches competitive performance with mainstream feature distillation methods. We promise to release our code and models to ensure reproducibility.
****

On Distillation of Guided Diffusion Models

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, Tim Salimans; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14297-14306

Classifier-free guided diffusion models have recently been shown to be highly effective at high-resolution image generation, and they have been widely used in large-scale diffusion frameworks including DALL*E 2, Stable Diffusion and Imagen. However, a downside of classifier-free guided diffusion models is that they are computationally expensive at inference time since they require evaluating two diffusion models, a class-conditional model and an unconditional model, tens to hundreds of times. To deal with this limitation, we propose an approach to distilling classifier-free guided diffusion models into models that are fast to sample from: Given a pre-trained classifier-free guided model, we first learn a single model to match the output of the combined conditional and unconditional models, and then we progressively distill that model to a diffusion model that requires much fewer sampling steps. For standard diffusion models trained on the pixel-space, our approach is able to generate images visually comparable to that of the original model using as few as 4 sampling steps on ImageNet 64x64 and CIFAR-10, achieving FID/IS scores comparable to that of the original model while being up to 256 times faster to sample from. For diffusion models trained on the latent-space (e.g., Stable Diffusion), our approach is able to generate high-fidelity images using as few as 1 to 4 denoising steps, accelerating inference by at least 10-fold compared to existing methods on ImageNet 256x256 and LAION datasets. We further demonstrate the effectiveness of our approach on text-guided image editing and inpainting, where our distilled model is able to generate high-quality results using as few as 2-4 denoising steps.
****

Dual-Path Adaptation From Image to Video Transformers

Jungin Park, Jiyoung Lee, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2203-2213

In this paper, we efficiently transfer the surpassing representation power of the vision foundation models, such as ViT and Swin, for video understanding with only a few trainable parameters. Previous adaptation methods have simultaneously considered spatial and temporal modeling with a unified learnable module but still suffered from fully leveraging the representative capabilities of image transformers. We argue that the popular dual-path (two-stream) architecture in video models can mitigate this problem. We propose a novel DUALPATH adaptation separated into spatial and temporal adaptation paths, where a lightweight bottleneck adapter is employed in each transformer block. Especially for temporal dynamic modeling, we incorporate consecutive frames into a grid-like frameset to precisely imitate vision transformers' capability that extrapolates relationships between tokens. In addition, we extensively investigate the multiple baselines from a unified perspective in video understanding and compare them with DUALPATH. Experimental results on four action recognition benchmarks prove that pretrained image transformers with DUALPATH can be effectively generalized beyond the data domain.
****

Towards Better Decision Forests: Forest Alternating Optimization
Miguel Á. Carreira-Perpiñán, Magzhan Gabidolla, Arman Zharmagambetov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7589-7598
Decision forests are among the most accurate models in machine learning. This is remarkable given that the way they are trained is highly heuristic: neither the individual trees nor the overall forest optimize any well-defined loss. While diversity mechanisms such as bagging or boosting have been until now critical in the success of forests, we think that a better optimization should lead to better forests---ideally eliminating any need for an ensembling heuristic. However, unlike for most other models, such as neural networks, optimizing forests or trees is not easy, because they define a non-differentiable function. We show, for the first time, that it is possible to learn a forest by optimizing a desirable loss and regularization jointly over all its trees and parameters. Our algorithm, Forest Alternating Optimization, is based on defining a forest as a parametric model with a fixed number of trees and structure (rather than adding trees indefinitely as in bagging or boosting). It then iteratively updates each tree in alternation so that the objective function decreases monotonically. The algorithm is so effective at optimizing that it easily overfits, but this can be corrected by averaging. The result is a forest that consistently exceeds the accuracy of the state-of-the-art while using fewer, smaller trees.
****
DA Wand: Distortion-Aware Selection Using Neural Mesh Parameterization
Richard Liu, Noam Aigerman, Vladimir G. Kim, Rana Hanocka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16739-16749
We present a neural technique for learning to select a local sub-region around a point which can be used for mesh parameterization. The motivation for our framework is driven by interactive workflows used for decaling, texturing, or painting on surfaces. Our key idea to to learn a local parameterization in a data-driven manner, using a novel differentiable parameterization layer within a neural network framework. We train a segmentation network to select 3D regions that are parameterized into 2D and penalized by the resulting distortion, giving rise to segmentations which are distortion-aware. Following training, a user can use our system to interactively select a point on the mesh and obtain a large, meaningful region around the selection which induces a low-distortion parameterization. Our code and project page are publicly available.
****
Disentangled Representation Learning for Unsupervised Neural Quantization
Haechan Noh, Sangeek Hyun, Woojin Jeong, Hanshin Lim, Jae-Pil Heo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12001-12010
The inverted index is a widely used data structure to avoid the infeasible exhaustive search. It accelerates retrieval significantly by splitting the database into multiple disjoint sets and restricts distance computation to a small fraction of the database. Moreover, it even improves search quality by allowing quantizers to exploit the compact distribution of residual vector space. However, we firstly point out a problem that an existing deep learning-based quantizer hardly benefits from the residual vector space, unlike conventional shallow quantizers. To cope with this problem, we introduce a novel disentangled representation learning for unsupervised neural quantization. Similar to the concept of residual vector space, the proposed method enables more compact latent space by disentangling information of the inverted index from the vectors. Experimental results on large-scale datasets confirm that our method outperforms the state-of-the-art retrieval systems by a large margin.
****
Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding
Chaolei Tan, Zihang Lin, Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18973-18982

Video Paragraph Grounding (VPG) is an essential yet challenging task in vision-l
anguage understanding, which aims to jointly localize multiple events from an un
trimmed video with a paragraph query description. One of the critical challenges
 in addressing this problem is to comprehend the complex semantic relations betw
een visual and textual modalities. Previous methods focus on modeling the contex
tual information between the video and text from a single-level perspective (i.e
., the sentence level), ignoring rich visual-textual correspondence relations at
 different semantic levels, e.g., the video-word and video-paragraph corresponde
nce. To this end, we propose a novel Hierarchical Semantic Correspondence Networ
k (HSCNet), which explores multi-level visual-textual correspondence by learning
 hierarchical semantic alignment and utilizes dense supervision by grounding div
erse levels of queries. Specifically, we develop a hierarchical encoder that enc
odes the multi-modal inputs into semantics-aligned representations at different
levels. To exploit the hierarchical semantic correspondence learned in the encod
er for multi-level supervision, we further design a hierarchical decoder that pr
ogressively performs finer grounding for lower-level queries conditioned on high
er-level semantics. Extensive experiments demonstrate the effectiveness of HSCNe
t and our method significantly outstrips the state-of-the-arts on two challengin
g benchmarks, i.e., ActivityNet-Captions and TACoS.
****
Temporal Attention Unit: Towards Efficient Spatiotemporal Predictive Learning
Cheng Tan, Zhangyang Gao, Lirong Wu, Yongjie Xu, Jun Xia, Siyuan Li, Stan Z. Li;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2023, pp. 18770-18782
Spatiotemporal predictive learning aims to generate future frames by learning fr
om historical frames. In this paper, we investigate existing methods and present
 a general framework of spatiotemporal predictive learning, in which the spatial
 encoder and decoder capture intra-frame features and the middle temporal module
 catches inter-frame correlations. While the mainstream methods employ recurrent
 units to capture long-term temporal dependencies, they suffer from low computat
ional efficiency due to their unparallelizable architectures. To parallelize the
 temporal module, we propose the Temporal Attention Unit (TAU), which decomposes
 temporal attention into intra-frame statical attention and inter-frame dynamica
l attention. Moreover, while the mean squared error loss focuses on intra-frame
errors, we introduce a novel differential divergence regularization to take inte
r-frame variations into account. Extensive experiments demonstrate that the prop
osed method enables the derived model to achieve competitive performance on vari
ous spatiotemporal prediction benchmarks.
****
Zero-Shot Pose Transfer for Unrigged Stylized 3D Characters
Jiashun Wang, Xueting Li, Sifei Liu, Shalini De Mello, Orazio Gallo, Xiaolong Wa
ng, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2023, pp. 8704-8714
Transferring the pose of a reference avatar to stylized 3D characters of various
 shapes is a fundamental task in computer graphics. Existing methods either requ
ire the stylized characters to be rigged, or they use the stylized character in
the desired pose as ground truth at training. We present a zero-shot approach th
at requires only the widely available deformed non-stylized avatars in training,
 and deforms stylized characters of significantly different shapes at inference.
 Classical methods achieve strong generalization by deforming the mesh at the tr
iangle level, but this requires labelled correspondences. We leverage the power
of local deformation, but without requiring explicit correspondence labels. We i
ntroduce a semi-supervised shape-understanding module to bypass the need for exp
licit correspondences at test time, and an implicit pose deformation module that
 deforms individual surface points to match the target pose. Furthermore, to enc
ourage realistic and accurate deformation of stylized characters, we introduce a
n efficient volume-based test-time training procedure. Because it does not need
rigging, nor the deformed stylized character at training time, our model general
izes to categories with scarce annotation, such as stylized quadrupeds. Extensiv
e experiments demonstrate the effectiveness of the proposed method compared to t

he state-of-the-art approaches trained with comparable or more supervision. Our project page is available at https://jiashunwang.github.io/ZPT
****
Listening Human Behavior: 3D Human Pose Estimation With Acoustic Signals
Yuto Shibata, Yutaka Kawashima, Mariko Isogawa, Go Irie, Akisato Kimura, Yoshimitsu Aoki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13323-13332
Given only acoustic signals without any high-level information, such as voices or sounds of scenes/actions, how much can we infer about the behavior of humans? Unlike existing methods, which suffer from privacy issues because they use signals that include human speech or the sounds of specific actions, we explore how low-level acoustic signals can provide enough clues to estimate 3D human poses by active acoustic sensing with a single pair of microphones and loudspeakers (see Fig. 1). This is a challenging task since sound is much more diffractive than other signals and therefore covers up the shape of objects in a scene. Accordingly, we introduce a framework that encodes multichannel audio features into 3D human poses. Aiming to capture subtle sound changes to reveal detailed pose information, we explicitly extract phase features from the acoustic signals together with typical spectrum features and feed them into our human pose estimation network. Also, we show that reflected or diffracted sounds are easily influenced by subjects' physique differences e.g., height and muscularity, which deteriorates prediction accuracy. We reduce these gaps by using a subject discriminator to improve accuracy. Our experiments suggest that with the use of only low-dimensional acoustic information, our method outperforms baseline methods. The datasets and codes used in this project will be publicly available.
****
Meta-Learning With a Geometry-Adaptive Preconditioner
Suhyun Kang, Duhun Hwang, Moonjung Eo, Taesup Kim, Wonjong Rhee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16080-16090
Model-agnostic meta-learning (MAML) is one of the most successful meta-learning algorithms. It has a bi-level optimization structure where the outer-loop process learns a shared initialization and the inner-loop process optimizes task-specific weights. Although MAML relies on the standard gradient descent in the inner-loop, recent studies have shown that controlling the inner-loop's gradient descent with a meta-learned preconditioner can be beneficial. Existing preconditioners, however, cannot simultaneously adapt in a task-specific and path-dependent way. Additionally, they do not satisfy the Riemannian metric condition, which can enable the steepest descent learning with preconditioned gradient. In this study, we propose Geometry-Adaptive Preconditioned gradient descent (GAP) that can overcome the limitations in MAML; GAP can efficiently meta-learn a preconditioner that is dependent on task-specific parameters, and its preconditioner can be shown to be a Riemannian metric. Thanks to the two properties, the geometry-adaptive preconditioner is effective for improving the inner-loop optimization. Experiment results show that GAP outperforms the state-of-the-art MAML family and preconditioned gradient descent-MAML (PGD-MAML) family in a variety of few-shot learning tasks. Code is available at: https://github.com/Suhyun777/CVPR23-GAP.
****
Dynamic Graph Enhanced Contrastive Learning for Chest X-Ray Report Generation
Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, Xiaojun Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3334-3343
Automatic radiology reporting has great clinical potential to relieve radiologists from heavy workloads and improve diagnosis interpretation. Recently, researchers have enhanced data-driven neural networks with medical knowledge graphs to eliminate the severe visual and textual bias in this task. The structures of such graphs are exploited by using the clinical dependencies formed by the disease topic tags via general knowledge and usually do not update during the training process. Consequently, the fixed graphs can not guarantee the most appropriate scope of knowledge and limit the effectiveness. To address the limitation, we propo

se a knowledge graph with Dynamic structure and nodes to facilitate chest X-ray report generation with Contrastive Learning, named DCL. In detail, the fundamental structure of our graph is pre-constructed from general knowledge. Then we explore specific knowledge extracted from the retrieved reports to add additional nodes or redefine their relations in a bottom-up manner. Each image feature is integrated with its very own updated graph before being fed into the decoder module for report generation. Finally, this paper introduces Image-Report Contrastive and Image-Report Matching losses to better represent visual features and textual information. Evaluated on IU-Xray and MIMIC-CXR datasets, our DCL outperforms previous state-of-the-art models on these two benchmarks.
****
BiCro: Noisy Correspondence Rectification for Multi-Modality Data via Bi-Directional Cross-Modal Similarity Consistency
Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, Min Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19883-19892
As one of the most fundamental techniques in multimodal learning, cross-modal matching aims to project various sensory modalities into a shared feature space. To achieve this, massive and correctly aligned data pairs are required for model training. However, unlike unimodal datasets, multimodal datasets are extremely harder to collect and annotate precisely. As an alternative, the co-occurred data pairs (e.g., image-text pairs) collected from the Internet have been widely exploited in the area. Unfortunately, the cheaply collected dataset unavoidably contains many mismatched data pairs, which have been proven to be harmful to the model's performance. To address this, we propose a general framework called BiCro (Bidirectional Cross-modal similarity consistency), which can be easily integrated into existing cross-modal matching models and improve their robustness against noisy data. Specifically, BiCro aims to estimate soft labels for noisy data pairs to reflect their true correspondence degree. The basic idea of BiCro is motivated by that -- taking image-text matching as an example -- similar images should have similar textual descriptions and vice versa. Then the consistency of these two similarities can be recast as the estimated soft labels to train the matching model. The experiments on three popular cross-modal matching datasets demonstrate that our method significantly improves the noise-robustness of various matching models, and surpass the state-of-the-art by a clear margin.
****
Transfer Knowledge From Head to Tail: Uncertainty Calibration Under Long-Tailed Distribution
Jiahao Chen, Bing Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19978-19987
How to estimate the uncertainty of a given model is a crucial problem. Current calibration techniques treat different classes equally and thus implicitly assume that the distribution of training data is balanced, but ignore the fact that real-world data often follows a long-tailed distribution. In this paper, we explore the problem of calibrating the model trained from a long-tailed distribution. Due to the difference between the imbalanced training distribution and balanced test distribution, existing calibration methods such as temperature scaling cannot generalize well to this problem. Specific calibration methods for domain adaptation are also not applicable because they rely on unlabeled target domain instances which are not available. Models trained from a long-tailed distribution tend to be more overconfident to head classes. To this end, we propose a novel knowledge-transferring-based calibration method by estimating the importance weights for samples of tail classes to realize long-tailed calibration. Our method models the distribution of each class as a Gaussian distribution and views the source statistics of head classes as a prior to calibrate the target distributions of tail classes. We adaptively transfer knowledge from head classes to get the target probability density of tail classes. The importance weight is estimated by the ratio of the target probability density over the source probability density. Extensive experiments on CIFAR-10-LT, MNIST-LT, CIFAR-100-LT, and ImageNet-LT datasets demonstrate the effectiveness of our method.

****
FrustumFormer: Adaptive Instance-Aware Resampling for Multi-View 3D Detection
Yuqi Wang, Yuntao Chen, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5096-5105
The transformation of features from 2D perspective space to 3D space is essential to multi-view 3D object detection. Recent approaches mainly focus on the design of view transformation, either pixel-wisely lifting perspective view features into 3D space with estimated depth or grid-wisely constructing BEV features via 3D projection, treating all pixels or grids equally. However, choosing what to transform is also important but has rarely been discussed before. The pixels of a moving car are more informative than the pixels of the sky. To fully utilize the information contained in images, the view transformation should be able to adapt to different image regions according to their contents. In this paper, we propose a novel framework named FrustumFormer, which pays more attention to the features in instance regions via adaptive instance-aware resampling. Specifically, the model obtains instance frustums on the bird's eye view by leveraging image view object proposals. An adaptive occupancy mask within the instance frustum is learned to refine the instance location. Moreover, the temporal frustum intersection could further reduce the localization uncertainty of objects. Comprehensive experiments on the nuScenes dataset demonstrate the effectiveness of FrustumFormer, and we achieve a new state-of-the-art performance on the benchmark. Codes and models will be made available at https://github.com/Robertwyq/Frustum.
****
Global Vision Transformer Pruning With Hessian-Aware Saliency
Huanrui Yang, Hongxu Yin, Maying Shen, Pavlo Molchanov, Hai Li, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18547-18557
Transformers yield state-of-the-art results across many tasks. However, their heuristically designed architecture impose huge computational costs during inference. This work aims on challenging the common design philosophy of the Vision Transformer (ViT) model with uniform dimension across all the stacked blocks in a model stage, where we redistribute the parameters both across transformer blocks and between different structures within the block via the first systematic attempt on global structural pruning. Dealing with diverse ViT structural components, we derive a novel Hessian-based structural pruning criteria comparable across all layers and structures, with latency-aware regularization for direct latency reduction. Performing iterative pruning on the DeiT-Base model leads to a new architecture family called NViT (Novel ViT), with a novel parameter redistribution that utilizes parameters more efficiently. On ImageNet-1K, NViT-Base achieves a 2.6x FLOPs reduction, 5.1x parameter reduction, and 1.9x run-time speedup over the DeiT-Base model in a near lossless manner. Smaller NViT variants achieve more than 1% accuracy gain at the same throughput of the DeiT Small/Tiny variants, as well as a lossless 3.3x parameter reduction over the SWIN-Small model. These results outperform prior art by a large margin. Further analysis is provided on the parameter redistribution insight of NViT, where we show the high prunability of ViT models, distinct sensitivity within ViT block, and unique parameter distribution trend across stacked ViT blocks. Our insights provide viability for a simple yet effective parameter redistribution rule towards more efficient ViTs for off-the-shelf performance boost.
****
Class-Conditional Sharpness-Aware Minimization for Deep Long-Tailed Recognition
Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, Wei Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3499-3509
It's widely acknowledged that deep learning models with flatter minima in its loss landscape tend to generalize better. However, such property is under-explored in deep long-tailed recognition (DLTR), a practical problem where the model is required to generalize equally well across all classes when trained on highly imbalanced label distribution. In this paper, through empirical observations, we argue that sharp minima are in fact prevalent in deep longtailed models, whereas

naive integration of existing flattening operations into long-tailed learning al gorithms brings little improvement. Instead, we propose an effective twostage sh arpness-aware optimization approach based on the decoupling paradigm in DLTR. In the first stage, both the feature extractor and classifier are trained under pa rameter perturbations at a class-conditioned scale, which is theoretically motiv ated by the characteristic radius of flat minima under the PAC-Bayesian framewor k. In the second stage, we generate adversarial features with classbalanced samp ling to further robustify the classifier with the backbone frozen. Extensive exp eriments on multiple longtailed visual recognition benchmarks show that, our pro posed Class-Conditional Sharpness-Aware Minimization (CC-SAM), achieves competit ive performance compared to the state-of-the-arts. Code is available at https:// github.com/zzpustc/CC-SAM.
****

ScarceNet: Animal Pose Estimation With Scarce Annotations
Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17174-17183
Animal pose estimation is an important but under-explored task due to the lack o f labeled data. In this paper, we tackle the task of animal pose estimation with scarce annotations, where only a small set of labeled data and unlabeled images are available. At the core of the solution to this problem setting is the use o f the unlabeled data to compensate for the lack of well-labeled animal pose data . To this end, we propose the ScarceNet, a pseudo label-based approach to genera te artificial labels for the unlabeled images. The pseudo labels, which are gene rated with a model trained with the small set of labeled images, are generally n oisy and can hurt the performance when directly used for training. To solve this problem, we first use a small-loss trick to select reliable pseudo labels. Alth ough effective, the selection process is improvident since numerous high-loss sa mples are left unused. We further propose to identify reusable samples from the high-loss samples based on an agreement check. Pseudo labels are re-generated to provide supervision for those reusable samples. Lastly, we introduce a student- teacher framework to enforce a consistency constraint since there are still samp les that are neither reliable nor reusable. By combining the reliable pseudo lab el selection with the reusable sample re-labeling and the consistency constraint , we can make full use of the unlabeled data. We evaluate our approach on the ch allenging AP-10K dataset, where our approach outperforms existing semi-supervise d approaches by a large margin. We also test on the TigDog dataset, where our ap proach can achieve better performance than domain adaptation based approaches wh en only very few annotations are available. Our code is available at the project website.
****

OmniCity: Omnipotent City Understanding With Multi-Level and Multi-View Images
Weijia Li, Yawen Lai, Linning Xu, Yuanbo Xiangli, Jinhua Yu, Conghui He, Gui-Son g Xia, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17397-17407
This paper presents OmniCity, a new dataset for omnipotent city understanding fr om multi-level and multi-view images. More precisely, OmniCity contains multi-vi ew satellite images as well as street-level panorama and mono-view images, const ituting over 100K pixel-wise annotated images that are well-aligned and collecte d from 25K geo-locations in New York City. To alleviate the substantial pixel-wi se annotation efforts, we propose an efficient street-view image annotation pipe line that leverages the existing label maps of satellite view and the transforma tion relations between different views (satellite, panorama, and mono-view). Wit h the new OmniCity dataset, we provide benchmarks for a variety of tasks includi ng building footprint extraction, height estimation, and building plane/instance /fine-grained segmentation. Compared with existing multi-level and multi-view be nchmarks, OmniCity contains a larger number of images with richer annotation typ es and more views, provides more benchmark results of state-of-the-art models, a nd introduces a new task for fine-grained building instance segmentation on stre et-level panorama images. Moreover, OmniCity provides new problem settings for e xisting tasks, such as cross-view image matching, synthesis, segmentation, detec

tion, etc., and facilitates the developing of new methods for large-scale city understanding, reconstruction, and simulation. The OmniCity dataset as well as the benchmarks will be released at https://city-super.github.io/omnicity/.
****
Efficient On-Device Training via Gradient Filtering
Yuedong Yang, Guihong Li, Radu Marculescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3811-3820
Despite its importance for federated learning, continuous learning and many other applications, on-device training remains an open problem for EdgeAI. The problem stems from the large number of operations (e.g., floating point multiplications and additions) and memory consumption required during training by the back-propagation algorithm. Consequently, in this paper, we propose a new gradient filtering approach which enables on-device CNN model training. More precisely, our approach creates a special structure with fewer unique elements in the gradient map, thus significantly reducing the computational complexity and memory consumption of back propagation during training. Extensive experiments on image classification and semantic segmentation with multiple CNN models (e.g., MobileNet, DeepLabV3, UPerNet) and devices (e.g., Raspberry Pi and Jetson Nano) demonstrate the effectiveness and wide applicability of our approach. For example, compared to SOTA, we achieve up to 19x speedup and 77.1% memory savings on ImageNet classification with only 0.1% accuracy loss. Finally, our method is easy to implement and deploy; over 20x speedup and 90% energy savings have been observed compared to highly optimized baselines in MKLDNN and CUDNN on NVIDIA Jetson Nano. Consequently, our approach opens up a new direction of research with a huge potential for on-device training.
****
SViTT: Temporal Learning of Sparse Video-Text Transformers
Yi Li, Kyle Min, Subarna Tripathi, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18919-18929
Do video-text transformers learn to model temporal relationships across frames? Despite their immense capacity and the abundance of multimodal training data, recent work has revealed the strong tendency of video-text models towards frame-based spatial representations, while temporal reasoning remains largely unsolved. In this work, we identify several key challenges in temporal learning of video-text transformers: the spatiotemporal trade-off from limited network size; the curse of dimensionality for multi-frame modeling; and the diminishing returns of semantic information by extending clip length. Guided by these findings, we propose SViTT, a sparse video-text architecture that performs multi-frame reasoning with significantly lower cost than naive transformers with dense attention. Analogous to graph-based networks, SViTT employs two forms of sparsity: edge sparsity that limits the query-key communications between tokens in self-attention, and node sparsity that discards uninformative visual tokens. Trained with a curriculum which increases model sparsity with the clip length, SViTT outperforms dense transformer baselines on multiple video-text retrieval and question answering benchmarks, with a fraction of computational cost. Project page: http://svcl.ucsd.edu/projects/svitt.
****
NeuralDome: A Neural Modeling Pipeline on Multi-View Human-Object Interactions
Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, Jingya Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8834-8845
Humans constantly interact with objects in daily life tasks. Capturing such processes and subsequently conducting visual inferences from a fixed viewpoint suffers from occlusions, shape and texture ambiguities, motions, etc. To mitigate the problem, it is essential to build a training dataset that captures free-viewpoint interactions. We construct a dense multi-view dome to acquire a complex human object interaction dataset, named HODome, that consists of 71M frames on 10 subjects interacting with 23 objects. To process the HODome dataset, we develop NeuralDome, a layer-wise neural processing pipeline tailored for multi-view video

inputs to conduct accurate tracking, geometry reconstruction and free-view rendering, for both human subjects and objects. Extensive experiments on the HODome dataset demonstrate the effectiveness of NeuralDome on a variety of inference, modeling, and rendering tasks. Both the dataset and the NeuralDome tools will be disseminated to the community for further development, which can be found at https://juzezhang.github.io/NeuralDome
****

3D Human Mesh Estimation From Virtual Markers
Xiaoxuan Ma, Jiajun Su, Chunyu Wang, Wentao Zhu, Yizhou Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 534-543
Inspired by the success of volumetric 3D pose estimation, some recent human mesh estimators propose to estimate 3D skeletons as intermediate representations, from which, the dense 3D meshes are regressed by exploiting the mesh topology. However, body shape information is lost in extracting skeletons, leading to mediocre performance. The advanced motion capture systems solve the problem by placing dense physical markers on the body surface, which allows to extract realistic meshes from their non-rigid motions. However, they cannot be applied to wild images without markers. In this work, we present an intermediate representation, named virtual markers, which learns 64 landmark keypoints on the body surface based on the large-scale mocap data in a generative style, mimicking the effects of physical markers. The virtual markers can be accurately detected from wild images and can reconstruct the intact meshes with realistic shapes by simple interpolation. Our approach outperforms the state-of-the-art methods on three datasets. In particular, it surpasses the existing methods by a notable margin on the SURREAL dataset, which has diverse body shapes. Code is available at https://github.com/ShirleyMaxx/VirtualMarker.
****

CUDA: Convolution-Based Unlearnable Datasets
Vinu Sankar Sadasivan, Mahdi Soltanolkotabi, Soheil Feizi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3862-3871
Large-scale training of modern deep learning models heavily relies on publicly available data on the web. This potentially unauthorized usage of online data leads to concerns regarding data privacy. Recent works aim to make unlearnable data for deep learning models by adding small, specially designed noises to tackle this issue. However, these methods are vulnerable to adversarial training (AT) and/or are computationally heavy. In this work, we propose a novel, model-free, Convolution-based Unlearnable DAtaset (CUDA) generation technique. CUDA is generated using controlled class-wise convolutions with filters that are randomly generated via a private key. CUDA encourages the network to learn the relation between filters and labels rather than informative features for classifying the clean data. We develop some theoretical analysis demonstrating that CUDA can successfully poison Gaussian mixture data by reducing the clean data performance of the optimal Bayes classifier. We also empirically demonstrate the effectiveness of CUDA with various datasets (CIFAR-10, CIFAR-100, ImageNet-100, and Tiny-ImageNet), and architectures (ResNet-18, VGG-16, Wide ResNet-34-10, DenseNet-121, DeIT, EfficientNetV2-S, and MobileNetV2). Our experiments show that CUDA is robust to various data augmentations and training approaches such as smoothing, AT with different budgets, transfer learning, and fine-tuning. For instance, training a ResNet-18 on ImageNet-100 CUDA achieves only 8.96%, 40.08%, and 20.58% clean test accuracies with empirical risk minimization (ERM), L_infinity AT, and L_2 AT, respectively. Here, ERM on the clean training data achieves a clean test accuracy of 80.66%. CUDA exhibits unlearnability effect with ERM even when only a fraction of the training dataset is perturbed. Furthermore, we also show that CUDA is robust to adaptive defenses designed specifically to break it.
****

No One Left Behind: Improving the Worst Categories in Long-Tailed Learning
Yingxiao Du, Jianxin Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15804-15813

Unlike the case when using a balanced training dataset, the per-class recall (i.e., accuracy) of neural networks trained with an imbalanced dataset are known to vary a lot from category to category. The convention in long-tailed recognition is to manually split all categories into three subsets and report the average accuracy within each subset. We argue that under such an evaluation setting, some categories are inevitably sacrificed. On one hand, focusing on the average accuracy on a balanced test set incurs little penalty even if some worst performing categories have zero accuracy. On the other hand, classes in the "Few" subset do not necessarily perform worse than those in the "Many" or "Medium" subsets. We therefore advocate to focus more on improving the lowest recall among all categories and the harmonic mean of all recall values. Specifically, we propose a simple plug-in method that is applicable to a wide range of methods. By simply re-training the classifier of an existing pre-trained model with our proposed loss function and using an optional ensemble trick that combines the predictions of the two classifiers, we achieve a more uniform distribution of recall values across categories, which leads to a higher harmonic mean accuracy while the (arithmetic) average accuracy is still high. The effectiveness of our method is justified on widely used benchmark datasets.
****

Deep Fair Clustering via Maximizing and Minimizing Mutual Information: Theory, Algorithm and Metric

Pengxin Zeng, Yunfan Li, Peng Hu, Dezhong Peng, Jiancheng Lv, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23986-23995

Fair clustering aims to divide data into distinct clusters while preventing sensitive attributes (e.g., gender, race, RNA sequencing technique) from dominating the clustering. Although a number of works have been conducted and achieved huge success recently, most of them are heuristical, and there lacks a unified theory for algorithm design. In this work, we fill this blank by developing a mutual information theory for deep fair clustering and accordingly designing a novel algorithm, dubbed FCMI. In brief, through maximizing and minimizing mutual information, FCMI is designed to achieve four characteristics highly expected by deep fair clustering, i.e., compact, balanced, and fair clusters, as well as informative features. Besides the contributions to theory and algorithm, another contribution of this work is proposing a novel fair clustering metric built upon information theory as well. Unlike existing evaluation metrics, our metric measures the clustering quality and fairness as a whole instead of separate manner. To verify the effectiveness of the proposed FCMI, we conduct experiments on six benchmarks including a single-cell RNA-seq atlas compared with 11 state-of-the-art methods in terms of five metrics. The code could be accessed from https://pengxi.me.
****

MIANet: Aggregating Unbiased Instance and General Information for Few-Shot Semantic Segmentation

Yong Yang, Qiong Chen, Yuan Feng, Tianlin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7131-7140

Existing few-shot segmentation methods are based on the meta-learning strategy and extract instance knowledge from a support set and then apply the knowledge to segment target objects in a query set. However, the extracted knowledge is insufficient to cope with the variable intra-class differences since the knowledge is obtained from a few samples in the support set. To address the problem, we propose a multi-information aggregation network (MIANet) that effectively leverages the general knowledge, i.e., semantic word embeddings, and instance information for accurate segmentation. Specifically, in MIANet, a general information module (GIM) is proposed to extract a general class prototype from word embeddings as a supplement to instance information. To this end, we design a triplet loss that treats the general class prototype as an anchor and samples positive-negative pairs from local features in the support set. The calculated triplet loss can transfer semantic similarities among language identities from a word embedding space to a visual representation space. To alleviate the model biasing towards the seen training classes and to obtain multi-scale information, we then introduce a

 non-parametric hierarchical prior module (HPM) to generate unbiased instance-le
vel information via calculating the pixel-level similarity between the support a
nd query image features. Finally, an information fusion module (IFM) combines th
e general and instance information to make predictions for the query image. Exte
nsive experiments on PASCAL-5i and COCO-20i show that MIANet yields superior per
formance and set a new state-of-the-art. Code is available at github.com/Aldrich
2y/MIANet.
****
High Fidelity 3D Hand Shape Reconstruction via Scalable Graph Frequency Decompos
ition
Tianyu Luan, Yuanhao Zhai, Jingjing Meng, Zhong Li, Zhang Chen, Yi Xu, Junsong Y
uan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2023, pp. 16795-16804
Despite the impressive performance obtained by recent single-image hand modeling
 techniques, they lack the capability to capture sufficient details of the 3D ha
nd mesh. This deficiency greatly limits their applications when high fidelity ha
nd modeling is required, e.g., personalized hand modeling. To address this probl
em, we design a frequency split network to generate 3D hand mesh using different
 frequency bands in a coarse-to-fine manner. To capture high-frequency personali
zed details, we transform the 3D mesh into the frequency domain, and propose a n
ovel frequency decomposition loss to supervise each frequency component. By leve
raging such a coarse-to-fine scheme, hand details that correspond to the higher
frequency domain can be preserved. In addition, the proposed network is scalable
, and can stop the inference at any resolution level to accommodate different ha
rdwares with varying computational powers. To quantitatively evaluate the perfor
mance of our method in terms of recovering personalized shape details, we introd
uce a new evaluation metric named Mean Signal-to-Noise Ratio (MSNR) to measure t
he signal-to-noise ratio of each mesh frequency component. Extensive experiments
 demonstrate that our approach generates fine-grained details for high fidelity
3D hand reconstruction, and our evaluation metric is more effective for measurin
g mesh details compared with traditional metrics.
****
COT: Unsupervised Domain Adaptation With Clustering and Optimal Transport
Yang Liu, Zhipeng Zhou, Baigui Sun; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2023, pp. 19998-20007
Unsupervised domain adaptation (UDA) aims to transfer the knowledge from a label
ed source domain to an unlabeled target domain. Typically, to guarantee desirabl
e knowledge transfer, aligning the distribution between source and target domain
 from a global perspective is widely adopted in UDA. Recent researchers further
point out the importance of local-level alignment and propose to construct insta
nce-pair alignment by leveraging on Optimal Transport (OT) theory. However, exis
ting OT-based UDA approaches are limited to handling class imbalance challenges
and introduce a heavy computation overhead when considering a large-scale traini
ng situation. To cope with two aforementioned issues, we propose a Clustering-ba
sed Optimal Transport (COT) algorithm, which formulates the alignment procedure
as an Optimal Transport problem and constructs a mapping between clustering cent
ers in the source and target domain via an end-to-end manner. With this alignmen
t on clustering centers, our COT eliminates the negative effect caused by class
imbalance and reduces the computation cost simultaneously. Empirically, our COT
achieves state-of-the-art performance on several authoritative benchmark dataset
s.
****
Target-Referenced Reactive Grasping for Dynamic Objects
Jirong Liu, Ruo Zhang, Hao-Shu Fang, Minghao Gou, Hongjie Fang, Chenxi Wang, She
ng Xu, Hengxu Yan, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2023, pp. 8824-8833
Reactive grasping, which enables the robot to successfully grasp dynamic moving
objects, is of great interest in robotics. Current methods mainly focus on the t
emporal smoothness of the predicted grasp poses but few consider their semantic
consistency. Consequently, the predicted grasps are not guaranteed to fall on th

e same part of the same object, especially in cluttered scenes. In this paper, we propose to solve reactive grasping in a target-referenced setting by tracking through generated grasp spaces. Given a targeted grasp pose on an object and detected grasp poses in a new observation, our method is composed of two stages: 1) discovering grasp pose correspondences through an attentional graph neural network and selecting the one with the highest similarity with respect to the target pose; 2) refining the selected grasp poses based on target and historical information. We evaluate our method on a large-scale benchmark GraspNet-1Billion. We also collect 30 scenes of dynamic objects for testing. The results suggest that our method outperforms other representative methods. Furthermore, our real robot experiments achieve an average success rate of over 80 percent.
****
Learning To Exploit the Sequence-Specific Prior Knowledge for Image Processing Pipelines Optimization
Haina Qin, Longfei Han, Weihua Xiong, Juan Wang, Wentao Ma, Bing Li, Weiming Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22314-22323
The hardware image signal processing (ISP) pipeline is the intermediate layer between the imaging sensor and the downstream application, processing the sensor signal into an RGB image. The ISP is less programmable and consists of a series of processing modules. Each processing module handles a subtask and contains a set of tunable hyperparameters. A large number of hyperparameters form a complex mapping with the ISP output. The industry typically relies on manual and time-consuming hyperparameter tuning by image experts, biased towards human perception. Recently, several automatic ISP hyperparameter optimization methods using downstream evaluation metrics come into sight. However, existing methods for ISP tuning treat the high-dimensional parameter space as a global space for optimization and prediction all at once without inducing the structure knowledge of ISP. To this end, we propose a sequential ISP hyperparameter prediction framework that utilizes the sequential relationship within ISP modules and the similarity among parameters to guide the model sequence process. We validate the proposed method on object detection, image segmentation, and image quality tasks.
****
Complexity-Guided Slimmable Decoder for Efficient Deep Video Compression
Zhihao Hu, Dong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14358-14367
In this work, we propose the complexity-guided slimmable decoder (cgSlimDecoder) in combination with skip-adaptive entropy coding (SaEC) for efficient deep video compression. Specifically, given the target complexity constraints, in our cgSlimDecoder, we introduce a set of new channel width selection modules to automatically decide the optimal channel width of each slimmable convolution layer. By optimizing the complexity-rate-distortion related objective function to directly learn the parameters of the newly introduced channel width selection modules and other modules in the decoder, our cgSlimDecoder can automatically allocate the optimal numbers of parameters for different types of modules (e.g., motion/residual decoder and the motion compensation network) and simultaneously support multiple complexity levels by using a single learnt decoder instead of multiple decoders. In addition, our proposed SaEC can further accelerate the entropy decoding procedure in both motion and residual decoders by simply skipping the entropy coding process for the elements in the encoded feature maps that are already well-predicted by the hyperprior network. As demonstrated in our comprehensive experiments, our newly proposed methods cgSlimDecoder and SaEC are general and can be readily incorporated into three widely used deep video codecs (i.e., DVC, FVC and DCVC) to significantly improve their coding efficiency with negligible performance drop.
****
Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation
Ning Zhang, Francesco Nex, George Vosselman, Norman Kerle; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1

8537-18546
Self-supervised monocular depth estimation that does not require ground truth for training has attracted attention in recent years. It is of high interest to design lightweight but effective models so that they can be deployed on edge devices. Many existing architectures benefit from using heavier backbones at the expense of model sizes. This paper achieves comparable results with a lightweight architecture. Specifically, the efficient combination of CNNs and Transformers is investigated, and a hybrid architecture called Lite-Mono is presented. A Consecutive Dilated Convolutions (CDC) module and a Local-Global Features Interaction (LGFI) module are proposed. The former is used to extract rich multi-scale local features, and the latter takes advantage of the self-attention mechanism to encode long-range global information into the features. Experiments demonstrate that Lite-Mono outperforms Monodepth2 by a large margin in accuracy, with about 80% fewer trainable parameters. Our codes and models are available at https://github.com/noahzn/Lite-Mono.
****
MarginMatch: Improving Semi-Supervised Learning with Pseudo-Margins
Tiberiu Sosea, Cornelia Caragea; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15773-15782
We introduce MarginMatch, a new SSL approach combining consistency regularization and pseudo-labeling, with its main novelty arising from the use of unlabeled data training dynamics to measure pseudo-label quality. Instead of using only the model's confidence on an unlabeled example at an arbitrary iteration to decide if the example should be masked or not, MarginMatch also analyzes the behavior of the model on the pseudo-labeled examples as the training progresses, ensuring low fluctuations in the model's predictions from one iteration to another. MarginMatch brings substantial improvements on four vision benchmarks in low data regimes and on two large-scale datasets, emphasizing the importance of enforcing high-quality pseudo-labels. Notably, we obtain an improvement in error rate over the state-of-the-art of 3.25% on CIFAR-100 with only 25 examples per class and of 4.19% on STL-10 using as few as 4 examples per class.
****
Neural Scene Chronology
Haotong Lin, Qianqian Wang, Ruojin Cai, Sida Peng, Hadar Averbuch-Elor, Xiaowei Zhou, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20752-20761
In this work, we aim to reconstruct a time-varying 3D model, capable of rendering photo-realistic renderings with independent control of viewpoint, illumination, and time, from Internet photos of large-scale landmarks. The core challenges are twofold. First, different types of temporal changes, such as illumination and changes to the underlying scene itself (such as replacing one graffiti artwork with another) are entangled together in the imagery. Second, scene-level temporal changes are often discrete and sporadic over time, rather than continuous. To tackle these problems, we propose a new scene representation equipped with a novel temporal step function encoding method that can model discrete scene-level content changes as piece-wise constant functions over time. Specifically, we represent the scene as a space-time radiance field with a per-image illumination embedding, where temporally-varying scene changes are encoded using a set of learned step functions. To facilitate our task of chronology reconstruction from Internet imagery, we also collect a new dataset of four scenes that exhibit various changes over time. We demonstrate that our method exhibits state-of-the-art view synthesis results on this dataset, while achieving independent control of viewpoint, time, and illumination. Code and data are available at https://zju3dv.github.io/NeuSC/.
****
Starting From Non-Parametric Networks for 3D Point Cloud Analysis
Renrui Zhang, Liuhui Wang, Yali Wang, Peng Gao, Hongsheng Li, Jianbo Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5344-5353
We present a Non-parametric Network for 3D point cloud analysis, Point-NN, which

consists of purely non-learnable components: farthest point sampling (FPS), k-nearest neighbors (k-NN), and pooling operations, with trigonometric functions. Surprisingly, it performs well on various 3D tasks, requiring no parameters or training, and even surpasses existing fully trained models. Starting from this basic non-parametric model, we propose two extensions. First, Point-NN can serve as a base architectural framework to construct Parametric Networks by simply inserting linear layers on top. Given the superior non-parametric foundation, the derived Point-PN exhibits a high performance-efficiency trade-off with only a few learnable parameters. Second, Point-NN can be regarded as a plug-and-play module for the already trained 3D models during inference. Point-NN captures the complementary geometric knowledge and enhances existing methods for different 3D benchmarks without re-training. We hope our work may cast a light on the community for understanding 3D point clouds with non-parametric methods. Code is available at https://github.com/ZrrSkywalker/Point-NN.
****
Light Source Separation and Intrinsic Image Decomposition Under AC Illumination
Yusaku Yoshida, Ryo Kawahara, Takahiro Okabe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5735-5743
Artificial light sources are often powered by an electric grid, and then their intensities rapidly oscillate in response to the grid's alternating current (AC). Interestingly, the flickers of scene radiance values due to AC illumination are useful for extracting rich information on a scene of interest. In this paper, we show that the flickers due to AC illumination is useful for intrinsic image decomposition (IID). Our proposed method conducts the light source separation (LSS) followed by the IID under AC illumination. In particular, we reveal the ambiguity in the blind LSS via matrix factorization and the ambiguity in the IID assuming the Lambert model, and then show why and how those ambiguities can be resolved. We experimentally confirmed that our method can recover the colors of the light sources, the diffuse reflectance values, and the diffuse and specular intensities (shadings) under each of the light sources, and that the IID under AC illumination is effective for application to auto white balancing.
****
TIPI: Test Time Adaptation With Transformation Invariance
A. Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, Philip H.S. Torr; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24162-24171
When deploying a machine learning model to a new environment, we often encounter the distribution shift problem -- meaning the target data distribution is different from the model's training distribution. In this paper, we assume that labels are not provided for this new domain, and that we do not store the source data (e.g., for privacy reasons). It has been shown that even small shifts in the data distribution can affect the model's performance severely. Test Time Adaptation offers a means to combat this problem, as it allows the model to adapt during test time to the new data distribution, using only unlabeled test data batches. To achieve this, the predominant approach is to optimize a surrogate loss on the test-time unlabeled target data. In particular, minimizing the prediction's entropy on target samples has received much interest as it is task-agnostic and does not require altering the model's training phase (e.g., does not require adding a self-supervised task during training on the source domain). However, as the target data's batch size is often small in real-world scenarios (e.g., autonomous driving models process each few frames in real-time), we argue that this surrogate loss is not optimal since it often collapses with small batch sizes. To tackle this problem, in this paper, we propose to use an invariance regularizer as the surrogate loss during test-time adaptation, motivated by our theoretical results regarding the model's performance under input transformations. The resulting method (TIPI -- Test tIme adaPtation with transformation Invariance) is validated with extensive experiments in various benchmarks (Cifar10-C, Cifar100-C, ImageNet-C, DIGITS, and VisDA17). Remarkably, TIPI is robust against small batch sizes (as small as 2 in our experiments), and consistently outperforms TENT in all settings. Our code is released at https://github.com/atuannguyen/TIPI.

****
OTAvatar: One-Shot Talking Face Avatar With Controllable Tri-Plane Rendering
Zhiyuan Ma, Xiangyu Zhu, Guo-Jun Qi, Zhen Lei, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16901-16910
Controllability, generalizability and efficiency are the major objectives of constructing face avatars represented by neural implicit field. However, existing methods have not managed to accommodate the three requirements simultaneously. They either focus on static portraits, restricting the representation ability to a specific subject, or suffer from substantial computational cost, limiting their flexibility. In this paper, we propose One-shot Talking face Avatar (OTAvatar), which constructs face avatars by a generalized controllable tri-plane rendering solution so that each personalized avatar can be constructed from only one portrait as the reference. Specifically, OTAvatar first inverts a portrait image to a motion-free identity code. Second, the identity code and a motion code are utilized to modulate an efficient CNN to generate a tri-plane formulated volume, which encodes the subject in the desired motion. Finally, volume rendering is employed to generate an image in any view. The core of our solution is a novel decoupling-by-inverting strategy that disentangles identity and motion in the latent code via optimization-based inversion. Benefiting from the efficient tri-plane representation, we achieve controllable rendering of generalized face avatar at 35 FPS on A100. Experiments show promising performance of cross-identity reenactment on subjects out of the training set and better 3D consistency. The code is available at https://github.com/theEricMa/OTAvatar.
****
Beyond Appearance: A Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks
Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, Xiuyu Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15050-15061
Human-centric visual tasks have attracted increasing research attention due to their widespread applications. In this paper, we aim to learn a general human representation from massive unlabeled human images which can benefit downstream human-centric tasks to the maximum extent. We call this method SOLIDER, a Semantic cOntrollable seLf-supervIseD lEaRning framework. Unlike the existing self-supervised learning methods, prior knowledge from human images is utilized in SOLIDER to build pseudo semantic labels and import more semantic information into the learned representation. Meanwhile, we note that different downstream tasks always require different ratios of semantic information and appearance information. For example, human parsing requires more semantic information, while person re-identification needs more appearance information for identification purpose. So a single learned representation cannot fit for all requirements. To solve this problem, SOLIDER introduces a conditional network with a semantic controller. After the model is trained, users can send values to the controller to produce representations with different ratios of semantic information, which can fit different needs of downstream tasks. Finally, SOLIDER is verified on six downstream human-centric visual tasks. It outperforms state of the arts and builds new baselines for these tasks. The code is released in https://github.com/tinyvision/SOLIDER.
****
Large-Capacity and Flexible Video Steganography via Invertible Neural Network
Chong Mou, Youmin Xu, Jiechong Song, Chen Zhao, Bernard Ghanem, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22606-22615
Video steganography is the art of unobtrusively concealing secret data in a cover video and then recovering the secret data through a decoding protocol at the receiver end. Although several attempts have been made, most of them are limited to low-capacity and fixed steganography. To rectify these weaknesses, we propose a Large-capacity and Flexible Video Steganography Network (LF-VSN) in this paper. For large-capacity, we present a reversible pipeline to perform multiple videos hiding and recovering through a single invertible neural network (INN). Our m

ethod can hide/recover 7 secret videos in/from 1 cover video with promising perf ormance. For flexibility, we propose a key-controllable scheme, enabling differe nt receivers to recover particular secret videos from the same cover video throu gh specific keys. Moreover, we further improve the flexibility by proposing a sc alable strategy in multiple videos hiding, which can hide variable numbers of se cret videos in a cover video with a single model and a single training session. Extensive experiments demonstrate that with the significant improvement of the v ideo steganography performance, our proposed LF-VSN has high security, large hid ing capacity, and flexibility. The source code is available at https://github.co m/MC-E/LF-VSN.
****

CFA: Class-Wise Calibrated Fair Adversarial Training
Zeming Wei, Yifei Wang, Yiwen Guo, Yisen Wang; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8193-8201
Adversarial training has been widely acknowledged as the most effective method t o improve the adversarial robustness against adversarial examples for Deep Neura l Networks (DNNs). So far, most existing works focus on enhancing the overall mo del robustness, treating each class equally in both the training and testing pha ses. Although revealing the disparity in robustness among classes, few works try to make adversarial training fair at the class level without sacrificing overal l robustness. In this paper, we are the first to theoretically and empirically i nvestigate the preference of different classes for adversarial configurations, i ncluding perturbation margin, regularization, and weight averaging. Motivated by this, we further propose a Class-wise calibrated Fair Adversarial training fram ework, named CFA, which customizes specific training configurations for each cla ss automatically. Experiments on benchmark datasets demonstrate that our propose d CFA can improve both overall robustness and fairness notably over other state-of-the-art methods. Code is available at https://github.com/PKU-ML/CFA.
****

EVAL: Explainable Video Anomaly Localization
Ashish Singh, Michael J. Jones, Erik G. Learned-Miller; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1871 7-18726
We develop a novel framework for single-scene video anomaly localization that al lows for human-understandable reasons for the decisions the system makes. We fir st learn general representations of objects and their motions (using deep networ ks) and then use these representations to build a high-level, location-dependent model of any particular scene. This model can be used to detect anomalies in ne w videos of the same scene. Importantly, our approach is explainable -- our high -level appearance and motion features can provide human-understandable reasons f or why any part of a video is classified as normal or anomalous. We conduct expe riments on standard video anomaly detection datasets (Street Scene, CUHK Avenue, ShanghaiTech and UCSD Ped1, Ped2) and show significant improvements over the pr evious state-of-the-art. All of our code and extra datasets will be made publicl y available.
****

Position-Guided Text Prompt for Vision-Language Pre-Training
Jinpeng Wang, Pan Zhou, Mike Zheng Shou, Shuicheng Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2324 2-23251
Vision-Language Pre-Training (VLP) has shown promising capabilities to align ima ge and text pairs, facilitating a broad variety of cross-modal learning tasks. H owever, we observe that VLP models often lack the visual grounding/localization capability which is critical for many downstream tasks such as visual reasoning. In this work, we propose a novel Position-guided Text Prompt (PTP) paradigm to enhance the visual grounding ability of cross-modal models trained with VLP. Spe cifically, in the VLP phase, PTP divides the image into NxN blocks, and identifi es the objects in each block through the widely used object detector in VLP. It then reformulates the visual grounding task into a fill-in-the-blank problem giv en a PTP by encouraging the model to predict the objects in the given blocks or

regress the blocks of a given object, e.g. filling "P" or "O" in a PTP "The block P has a O". This mechanism improves the visual grounding capability of VLP models and thus helps them better handle various downstream tasks. By introducing PTP into several state-of-the-art VLP frameworks, we observe consistently significant improvements across representative cross-modal learning model architectures and several benchmarks, e.g. zero-shot Flickr30K Retrieval (+4.8 in average recall@1) for ViLT baseline, and COCO Captioning (+5.3 in CIDEr) for SOTA BLIP baseline. Moreover, PTP achieves comparable results with object-detector based methods, and much faster inference speed since PTP discards its object detector for inference while the later cannot. Our code and pre-trained weight will be released.
****
HOLODIFFUSION: Training a 3D Diffusion Model Using 2D Images
Animesh Karnewar, Andrea Vedaldi, David Novotny, Niloy J. Mitra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18423-18433
Diffusion models have emerged as the best approach for generative modeling of 2D images. Part of their success is due to the possibility of training them on millions if not billions of images with a stable learning objective. However, extending these models to 3D remains difficult for two reasons. First, finding a large quantity of 3D training data is much more complex than for 2D images. Second, while it is conceptually trivial to extend the models to operate on 3D rather than 2D grids, the associated cubic growth in memory and compute complexity makes this infeasible. We address the first challenge by introducing a new diffusion setup that can be trained, end-to-end, with only posed 2D images for supervision; and the second challenge by proposing an image formation model that decouples model memory from spatial memory. We evaluate our method on real-world data, using the CO3D dataset which has not been used to train 3D generative models before. We show that our diffusion models are scalable, train robustly, and are competitive in terms of sample quality and fidelity to existing approaches for 3D generative modeling.
****
Stimulus Verification Is a Universal and Effective Sampler in Multi-Modal Human Trajectory Prediction
Jianhua Sun, Yuxuan Li, Liang Chai, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22014-22023
To comprehensively cover the uncertainty of the future, the common practice of multi-modal human trajectory prediction is to first generate a set/distribution of candidate future trajectories and then sample required numbers of trajectories from them as final predictions. Even though a large number of previous researches develop various strong models to predict candidate trajectories, how to effectively sample the final ones has not received much attention yet. In this paper, we propose stimulus verification, serving as a universal and effective sampling process to improve the multi-modal prediction capability, where stimulus refers to the factor in the observation that may affect the future movements such as social interaction and scene context. Stimulus verification introduces a probabilistic model, denoted as stimulus verifier, to verify the coherence between a predicted future trajectory and its corresponding stimulus. By highlighting prediction samples with better stimulus-coherence, stimulus verification ensures sampled trajectories plausible from the stimulus' point of view and therefore aids in better multi-modal prediction performance. We implement stimulus verification on five representative prediction frameworks and conduct exhaustive experiments on three widely-used benchmarks. Superior results demonstrate the effectiveness of our approach.
****
3D Human Pose Estimation With Spatio-Temporal Criss-Cross Attention
Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, Ting Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4790-4799
Recent transformer-based solutions have shown great success in 3D human pose est

imation. Nevertheless, to calculate the joint-to-joint affinity matrix, the comp utational cost has a quadratic growth with the increasing number of joints. Such drawback becomes even worse especially for pose estimation in a video sequence, which necessitates spatio-temporal correlation spanning over the entire video. In this paper, we facilitate the issue by decomposing correlation learning into space and time, and present a novel Spatio-Temporal Criss-cross attention (STC) block. Technically, STC first slices its input feature into two partitions evenl y along the channel dimension, followed by performing spatial and temporal atten tion respectively on each partition. STC then models the interactions between jo ints in an identical frame and joints in an identical trajectory simultaneously by concatenating the outputs from attention layers. On this basis, we devise STC Former by stacking multiple STC blocks and further integrate a new Structure-enh anced Positional Embedding (SPE) into STCFormer to take the structure of human b ody into consideration. The embedding function consists of two components: spati o-temporal convolution around neighboring joints to capture local structure, and part-aware embedding to indicate which part each joint belongs to. Extensive ex periments are conducted on Human3.6M and MPI-INF-3DHP benchmarks, and superior r esults are reported when comparing to the state-of-the-art approaches. More rema rkably, STCFormer achieves to-date the best published performance: 40.5mm P1 err or on the challenging Human3.6M dataset.
****
Plateau-Reduced Differentiable Path Tracing
Michael Fischer, Tobias Ritschel; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2023, pp. 4285-4294
Current differentiable renderers provide light transport gradients with respect to arbitrary scene parameters. However, the mere existence of these gradients do es not guarantee useful update steps in an optimization. Instead, inverse render ing might not converge due to inherent plateaus, i.e., regions of zero gradient, in the objective function. We propose to alleviate this by convolving the high-dimensional rendering function that maps scene parameters to images with an addi tional kernel that blurs the parameter space. We describe two Monte Carlo estima tors to compute plateau-free gradients efficiently, i.e., with low variance, and show that these translate into net-gains in optimization error and runtime perf ormance. Our approach is a straightforward extension to both black-box and diffe rentiable renderers and enables the successful optimization of problems with int ricate light transport, such as caustics or global illumination, that existing d ifferentiable path tracers do not converge on. Our code is at github.com/mfische r-ucl/prdpt.
****
LoGoNet: Towards Accurate 3D Object Detection With Local-to-Global Cross-Modal F usion
Xin Li, Tao Ma, Yuenan Hou, Botian Shi, Yuchen Yang, Youquan Liu, Xingjiao Wu, Q in Chen, Yikang Li, Yu Qiao, Liang He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17524-17534
LiDAR-camera fusion methods have shown impressive performance in 3D object detec tion. Recent advanced multi-modal methods mainly perform global fusion, where im age features and point cloud features are fused across the whole scene. Such pra ctice lacks fine-grained region-level information, yielding suboptimal fusion pe rformance. In this paper, we present the novel Local-to-Global fusion network (L oGoNet), which performs LiDAR-camera fusion at both local and global levels. Con cretely, the Global Fusion (GoF) of LoGoNet is built upon previous literature, w hile we exclusively use point centroids to more precisely represent the position of voxel features, thus achieving better cross-modal alignment. As to the Local Fusion (LoF), we first divide each proposal into uniform grids and then project these grid centers to the images. The image features around the projected grid points are sampled to be fused with position-decorated point cloud features, max imally utilizing the rich contextual information around the proposals. The Featu re Dynamic Aggregation (FDA) module is further proposed to achieve information i nteraction between these locally and globally fused features, thus producing mor e informative multi-modal features. Extensive experiments on both Waymo Open Dat

aset (WOD) and KITTI datasets show that LoGoNet outperforms all state-of-the-art 3D detection methods. Notably, LoGoNet ranks 1st on Waymo 3D object detection leaderboard and obtains 81.02 mAPH (L2) detection performance. It is noteworthy that, for the first time, the detection performance on three classes surpasses 80 APH (L2) simultaneously. Code will be available at https://github.com/sankin97/LoGoNet.
****
ScaleKD: Distilling Scale-Aware Knowledge in Small Object Detector
Yichen Zhu, Qiqi Zhou, Ning Liu, Zhiyuan Xu, Zhicai Ou, Xiaofeng Mou, Jian Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19723-19733
Despite the prominent success of general object detection, the performance and efficiency of Small Object Detection (SOD) are still unsatisfactory. Unlike existing works that struggle to balance the trade-off between inference speed and SOD performance, in this paper, we propose a novel Scale-aware Knowledge Distillation (ScaleKD), which transfers knowledge of a complex teacher model to a compact student model. We design two novel modules to boost the quality of knowledge transfer in distillation for SOD: 1) a scale-decoupled feature distillation module that disentangled teacher's feature representation into multi-scale embedding that enables explicit feature mimicking of the student model on small objects. 2) a cross-scale assistant to refine the noisy and uninformative bounding boxes prediction student models, which can mislead the student model and impair the efficacy of knowledge distillation. A multi-scale cross-attention layer is established to capture the multi-scale semantic information to improve the student model. We conduct experiments on COCO and VisDrone datasets with diverse types of models, i.e., two-stage and one-stage detectors, to evaluate our proposed method. Our ScaleKD achieves superior performance on general detection performance and obtains spectacular improvement regarding the SOD performance.
****
An Empirical Study of End-to-End Video-Language Transformers With Masked Visual Modeling
Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22898-22909
Masked visual modeling (MVM) has been recently proven effective for visual pre-training. While similar reconstructive objectives on video inputs (e.g., masked frame modeling) have been explored in video-language (VidL) pre-training, previous studies fail to find a truly effective MVM strategy that can largely benefit the downstream performance. In this work, we systematically examine the potential of MVM in the context of VidL learning. Specifically, we base our study on a fully end-to-end VIdeO-LanguagE Transformer (VIOLET), where the supervision from MVM training can be backpropagated to the video pixel space. In total, eight different reconstructive targets of MVM are explored, from low-level pixel values and oriented gradients to high-level depth maps, optical flow, discrete visual tokens, and latent visual features. We conduct comprehensive experiments and provide insights into the factors leading to effective MVM training, resulting in an enhanced model VIOLETv2. Empirically, we show VIOLETv2 pre-trained with MVM objective achieves notable improvements on 13 VidL benchmarks, ranging from video question answering, video captioning, to text-to-video retrieval.
****
Glocal Energy-Based Learning for Few-Shot Open-Set Recognition
Haoyu Wang, Guansong Pang, Peng Wang, Lei Zhang, Wei Wei, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7507-7516
Few-shot open-set recognition (FSOR) is a challenging task of great practical value. It aims to categorize a sample to one of the pre-defined, closed-set classes illustrated by few examples while being able to reject the sample from unknown classes. In this work, we approach the FSOR task by proposing a novel energy-based hybrid model. The model is composed of two branches, where a classification branch learns a metric to classify a sample to one of closed-set classes and the

energy branch explicitly estimates the open-set probability. To achieve holisti c detection of open-set samples, our model leverages both class-wise and pixel-w ise features to learn a glocal energy-based score, in which a global energy scor e is learned using the class-wise features, while a local energy score is learne d using the pixel-wise features. The model is enforced to assign large energy sc ores to samples that are deviated from the few-shot examples in either the class -wise features or the pixel-wise features, and to assign small energy scores oth erwise. Experiments on three standard FSOR datasets show the superior performanc e of our model.
****
Revisiting Temporal Modeling for CLIP-Based Image-to-Video Knowledge Transferrin g
Ruyang Liu, Jingjia Huang, Ge Li, Jiashi Feng, Xinglong Wu, Thomas H. Li; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2023, pp. 6555-6564
Image-text pretrained models, e.g., CLIP, have shown impressive general multi-mo dal knowledge learned from large-scale image-text data pairs, thus attracting in creasing attention for their potential to improve visual representation learning in the video domain. In this paper, based on the CLIP model, we revisit tempora l modeling in the context of image-to-video knowledge transferring, which is the key point for extending image-text pretrained models to the video domain. We fi nd that current temporal modeling mechanisms are tailored to either high-level s emantic-dominant tasks (e.g., retrieval) or low-level visual pattern-dominant ta sks (e.g., recognition), and fail to work on the two cases simultaneously. The k ey difficulty lies in modeling temporal dependency while taking advantage of bot h high-level and low-level knowledge in CLIP model. To tackle this problem, we p resent Spatial-Temporal Auxiliary Network (STAN) -- a simple and effective tempo ral modeling mechanism extending CLIP model to diverse video tasks. Specifically , to realize both low-level and high-level knowledge transferring, STAN adopts a branch structure with decomposed spatial-temporal modules that enable multi-lev el CLIP features to be spatial-temporally contextualized. We evaluate our method on two representative video tasks: Video-Text Retrieval and Video Recognition. Extensive experiments demonstrate the superiority of our model over the state-of -the-art methods on various datasets, including MSR-VTT, DiDeMo, LSMDC, MSVD, Ki netics-400, and Something-Something-V2. Codes will be available at https://githu b.com/farewellthree/STAN
****
MethaneMapper: Spectral Absorption Aware Hyperspectral Transformer for Methane D etection
Satish Kumar, Ivan Arevalo, ASM Iftekhar, B S Manjunath; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 176 09-17618
Methane (CH 4) is the chief contributor to global climate change. Recent Airbor ne Visible-Infrared Imaging Spectrometer-Next Generation (AVIRIS-NG) has been ve ry useful in quantitative mapping of methane emissions. Existing methods for ana lyzing this data are sensitive to local terrain conditions, often require manual inspection from domain experts, prone to significant error and hence are not sc alable. To address these challenges, we propose a novel end-to-end spectral abso rption wavelength aware transformer network, MethaneMapper, to detect and quanti fy the emissions. MethaneMapper introduces two novel modules that help to locate the most relevant methane plume regions in the spectral domain and uses them to localize these accurately. Thorough evaluation shows that MethaneMapper achieve s 0.63 mAP in detection and reduces the model size (by 5x) compared to the curre nt state of the art. In addition, we also introduce a large-scale dataset of met hane plume segmentation mask for over 1200 AVIRIS-NG flightlines from 2015-2022. It contains over 4000 methane plume sites. Our dataset will provide researchers the opportunity to develop and advance new methods for tackling this challengin g green-house gas detection problem with significant broader social impact. Data set and source code link.
****

Autonomous Manipulation Learning for Similar Deformable Objects via Only One Demonstration
Yu Ren, Ronghan Chen, Yang Cong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17069-17078
In comparison with most methods focusing on 3D rigid object recognition and manipulation, deformable objects are more common in our real life but attract less attention. Generally, most existing methods for deformable object manipulation suffer two issues, 1) Massive demonstration: repeating thousands of robot-object demonstrations for model training of one specific instance; 2) Poor generalization: inevitably re-training for transferring the learned skill to a similar/new instance from the same category. Therefore, we propose a category-level deformable 3D object manipulation framework, which could manipulate deformable 3D objects with only one demonstration and generalize the learned skills to new similar instances without re-training. Specifically, our proposed framework consists of two modules. The Nocs State Transform (NST) module transfers the observed point clouds of the target to a pre-defined unified pose state (i.e., Nocs state), which is the foundation for the category-level manipulation learning; the Neural Spatial Encoding (NSE) module generalizes the learned skill to novel instances by encoding the category-level spatial information to pursue the expected grasping point without re-training. The relative motion path is then planned to achieve autonomous manipulation. Both the simulated results via our Cap40 dataset and real robotic experiments justify the effectiveness of our framework.
****
Representation Learning for Visual Object Tracking by Masked Appearance Transfer
Haojie Zhao, Dong Wang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18696-18705
Visual representation plays an important role in visual object tracking. However, few works study the tracking-specified representation learning method. Most trackers directly use ImageNet pre-trained representations. In this paper, we propose masked appearance transfer, a simple but effective representation learning method for tracking, based on an encoder-decoder architecture. First, we encode the visual appearances of the template and search region jointly, and then we decode them separately. During decoding, the original search region image is reconstructed. However, for the template, we make the decoder reconstruct the target appearance within the search region. By this target appearance transfer, the tracking-specified representations are learned. We randomly mask out the inputs, thereby making the learned representations more discriminative. For sufficient evaluation, we design a simple and lightweight tracker that can evaluate the representation for both target localization and box regression. Extensive experiments show that the proposed method is effective, and the learned representations can enable the simple tracker to obtain state-of-the-art performance on six datasets.
****
EFEM: Equivariant Neural Field Expectation Maximization for 3D Object Segmentation Without Scene Supervision
Jiahui Lei, Congyue Deng, Karl Schmeckpeper, Leonidas Guibas, Kostas Daniilidis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4902-4912
We introduce Equivariant Neural Field Expectation Maximization (EFEM), a simple, effective, and robust geometric algorithm that can segment objects in 3D scenes without annotations or training on scenes. We achieve such unsupervised segmentation by exploiting single object shape priors. We make two novel steps in that direction. First, we introduce equivariant shape representations to this problem to eliminate the complexity induced by the variation in object configuration. Second, we propose a novel EM algorithm that can iteratively refine segmentation masks using the equivariant shape prior. We collect a novel real dataset Chairs and Mugs that contains various object configurations and novel scenes in order to verify the effectiveness and robustness of our method. Experimental results demonstrate that our method achieves consistent and robust performance across different scenes where the (weakly) supervised methods may fail. Code and data available at https://www.cis.upenn.edu/ leijh/projects/efem

****
Learning To Name Classes for Vision and Language Models
Sarah Parisot, Yongxin Yang, Steven McDonagh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23477-23486
Large scale vision and language models can achieve impressive zero-shot recognition performance by mapping class specific text queries to image content. Two distinct challenges that remain however, are high sensitivity to the choice of handcrafted class names that define queries, and the difficulty of adaptation to new, smaller datasets. Towards addressing these problems, we propose to leverage available data to learn, for each class, an optimal word embedding as a function of the visual content. By learning new word embeddings on an otherwise frozen model, we are able to retain zero-shot capabilities for new classes, easily adapt models to new datasets, and adjust potentially erroneous, non-descriptive or ambiguous class names. We show that our solution can easily be integrated in image classification and object detection pipelines, yields significant performance gains in multiple scenarios and provides insights into model biases and labelling errors.
****
ECON: Explicit Clothed Humans Optimized via Normal Integration
Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 512-523
The combination of deep learning, artist-curated scans, and Implicit Functions (IF), is enabling the creation of detailed, clothed, 3D humans from images. However, existing methods are far from perfect. IF-based methods recover free-form geometry, but produce disembodied limbs or degenerate shapes for novel poses or clothes. To increase robustness for these cases, existing work uses an explicit parametric body model to constrain surface reconstruction, but this limits the recovery of free-form surfaces such as loose clothing that deviates from the body. What we want is a method that combines the best properties of implicit representation and explicit body regularization. To this end, we make two key observations: (1) current networks are better at inferring detailed 2D maps than full-3D surfaces, and (2) a parametric model can be seen as a "canvas" for stitching together detailed surface patches. Based on these, our method, ECON, has three main steps: (1) It infers detailed 2D normal maps for the front and back side of a clothed person. (2) From these, it recovers 2.5D front and back surfaces, called d-BiNI, that are equally detailed, yet incomplete, and registers these w.r.t. each other with the help of a SMPL-X body mesh recovered from the image. (3) It "inpaints" the missing geometry between d-BiNI surfaces. If the face and hands are noisy, they can optionally be replaced with the ones of SMPL-X. As a result, ECON infers high-fidelity 3D humans even in loose clothes and challenging poses. This goes beyond previous methods, according to the quantitative evaluation on the CAPE and Renderpeople datasets. Perceptual studies also show that ECON's perceived realism is better by a large margin. Code and models are available for research purposes at econ.is.tue.mpg.de
****
Neural Fourier Filter Bank
Zhijie Wu, Yuhe Jin, Kwang Moo Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14153-14163
We present a novel method to provide efficient and highly detailed reconstructions. Inspired by wavelets, we learn a neural field that decompose the signal both spatially and frequency-wise. We follow the recent grid-based paradigm for spatial decomposition, but unlike existing work, encourage specific frequencies to be stored in each grid via Fourier features encodings. We then apply a multi-layer perceptron with sine activations, taking these Fourier encoded features in at appropriate layers so that higher-frequency components are accumulated on top of lower-frequency components sequentially, which we sum up to form the final output. We demonstrate that our method outperforms the state of the art regarding model compactness and convergence speed on multiple tasks: 2D image fitting, 3D shape reconstruction, and neural radiance fields. Our code is available at https:/

/github.com/ubc-vision/NFFB.
****
F2-NeRF: Fast Neural Radiance Field Training With Free Camera Trajectories
Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4150-4159
This paper presents a novel grid-based NeRF called F^2-NeRF (Fast-Free-NeRF) for novel view synthesis, which enables arbitrary input camera trajectories and only costs a few minutes for training. Existing fast grid-based NeRF training frameworks, like Instant-NGP, Plenoxels, DVGO, or TensoRF, are mainly designed for bounded scenes and rely on space warping to handle unbounded scenes. Existing two widely-used space-warping methods are only designed for the forward-facing trajectory or the 360deg object-centric trajectory but cannot process arbitrary trajectories. In this paper, we delve deep into the mechanism of space warping to handle unbounded scenes. Based on our analysis, we further propose a novel space-warping method called perspective warping, which allows us to handle arbitrary trajectories in the grid-based NeRF framework. Extensive experiments demonstrate that F^2-NeRF is able to use the same perspective warping to render high-quality images on two standard datasets and a new free trajectory dataset collected by us.
****
NeRFInvertor: High Fidelity NeRF-GAN Inversion for Single-Shot Real Image Animation
Yu Yin, Kamran Ghasedi, HsiangTao Wu, Jiaolong Yang, Xin Tong, Yun Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8539-8548
Nerf-based Generative models have shown impressive capacity in generating high-quality images with consistent 3D geometry. Despite successful synthesis of fake identity images randomly sampled from latent space, adopting these models for generating face images of real subjects is still a challenging task due to its so-called inversion issue. In this paper, we propose a universal method to surgically fine-tune these NeRF-GAN models in order to achieve high-fidelity animation of real subjects only by a single image. Given the optimized latent code for an out-of-domain real image, we employ 2D loss functions on the rendered image to reduce the identity gap. Furthermore, our method leverages explicit and implicit 3D regularizations using the in-domain neighborhood samples around the optimized latent code to remove geometrical and visual artifacts. Our experiments confirm the effectiveness of our method in realistic, high-fidelity, and 3D consistent animation of real faces on multiple NeRF-GAN models across different datasets.
****
Learning To Detect and Segment for Open Vocabulary Object Detection
Tao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7051-7060
Open vocabulary object detection has been greately advanced by the recent development of vision-language pre-trained model, which helps recognizing the novel objects with only semantic categories. The prior works mainly focus on knowledge transferring to the object proposal classification and employ class-agnostic box and mask prediction. In this work, we propose CondHead, a principled dynamic network design to better generalize the box regression and mask segmentation for open vocabulary setting. The core idea is to conditionally parametrize the network heads on semantic embedding and thus the model is guided with class-specific knowledge to better detect novel categories. Specifically, CondHead is composed of two streams of network heads, the dynamically aggregated heads and dynamically generated heads. The former is instantiated with a set of static heads that are conditionally aggregated, these heads are optimized as experts and are expected to learn sophisticated prediction. The Latter is instantiated with dynamically generated parameters and encodes general class-specific information. With such conditional design, the detection model is bridged by the semantic embedding to offer strongly generalizable class-wise box and mask prediction. Our method brings significant improvement to the prior state-of-the-art open vocabulary object de

tection methods with very minor overhead, e.g., it surpasses a RegionClip model by 3.0 detection AP on novel categories, with only 1.1% more computation.
****
Disentangling Writer and Character Styles for Handwriting Generation
Gang Dai, Yifan Zhang, Qingfeng Wang, Qing Du, Zhuliang Yu, Zhuoman Liu, Shuangping Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5977-5986
Training machines to synthesize diverse handwritings is an intriguing task. Recently, RNN-based methods have been proposed to generate stylized online Chinese characters. However, these methods mainly focus on capturing a person's overall writing style, neglecting subtle style inconsistencies between characters written by the same person. For example, while a person's handwriting typically exhibits general uniformity (e.g., glyph slant and aspect ratios), there are still small style variations in finer details (e.g., stroke length and curvature) of characters. In light of this, we propose to disentangle the style representations at both writer and character levels from individual handwritings to synthesize realistic stylized online handwritten characters. Specifically, we present the style-disentangled Transformer (SDT), which employs two complementary contrastive objectives to extract the style commonalities of reference samples and capture the detailed style patterns of each sample, respectively. Extensive experiments on various language scripts demonstrate the effectiveness of SDT. Notably, our empirical findings reveal that the two learned style representations provide information at different frequency magnitudes, underscoring the importance of separate style extraction. Our source code is public at: https://github.com/dailenson/SDT.
****
Nighttime Smartphone Reflective Flare Removal Using Optical Center Symmetry Prior
Yuekun Dai, Yihang Luo, Shangchen Zhou, Chongyi Li, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20783-20791
Reflective flare is a phenomenon that occurs when light reflects inside lenses, causing bright spots or a "ghosting effect" in photos, which can impact their quality. Eliminating reflective flare is highly desirable but challenging. Many existing methods rely on manually designed features to detect these bright spots, but they often fail to identify reflective flares created by various types of light and may even mistakenly remove the light sources in scenarios with multiple light sources. To address these challenges, we propose an optical center symmetry prior, which suggests that the reflective flare and light source are always symmetrical around the lens's optical center. This prior helps to locate the reflective flare's proposal region more accurately and can be applied to most smartphone cameras. Building on this prior, we create the first reflective flare removal dataset called BracketFlare, which contains diverse and realistic reflective flare patterns. We use continuous bracketing to capture the reflective flare pattern in the underexposed image and combine it with a normally exposed image to synthesize a pair of flare-corrupted and flare-free images. With the dataset, neural networks can be trained to remove the reflective flares effectively. Extensive experiments demonstrate the effectiveness of our method on both synthetic and real-world datasets.
****
StyleSync: High-Fidelity Generalized and Personalized Lip Sync in Style-Based Generator
Jiazhi Guan, Zhanwang Zhang, Hang Zhou, Tianshu Hu, Kaisiyuan Wang, Dongliang He, Haocheng Feng, Jingtuo Liu, Errui Ding, Ziwei Liu, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1505-1515
Despite recent advances in syncing lip movements with any audio waves, current methods still struggle to balance generation quality and the model's generalization ability. Previous studies either require long-term data for training or produce a similar movement pattern on all subjects with low quality. In this paper, we propose StyleSync, an effective framework that enables high-fidelity lip synch

ronization. We identify that a style-based generator would sufficiently enable such a charming property on both one-shot and few-shot scenarios. Specifically, we design a mask-guided spatial information encoding module that preserves the details of the given face. The mouth shapes are accurately modified by audio through modulated convolutions. Moreover, our design also enables personalized lip-sync by introducing style space and generator refinement on only limited frames. Thus the identity and talking style of a target person could be accurately preserved. Extensive experiments demonstrate the effectiveness of our method in producing high-fidelity results on a variety of scenes.
****

Balanced Spherical Grid for Egocentric View Synthesis

Changwoon Choi, Sang Min Kim, Young Min Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16590-16599

We present EgoNeRF, a practical solution to reconstruct large-scale real-world environments for VR assets. Given a few seconds of casually captured 360 video, EgoNeRF can efficiently build neural radiance fields which enable high-quality rendering from novel viewpoints. Motivated by the recent acceleration of NeRF using feature grids, we adopt spherical coordinate instead of conventional Cartesian coordinate. Cartesian feature grid is inefficient to represent large-scale unbounded scenes because it has a spatially uniform resolution, regardless of distance from viewers. The spherical parameterization better aligns with the rays of egocentric images, and yet enables factorization for performance enhancement. However, the naive spherical grid suffers from irregularities at two poles, and also cannot represent unbounded scenes. To avoid singularities near poles, we combine two balanced grids, which results in a quasi-uniform angular grid. We also partition the radial grid exponentially and place an environment map at infinity to represent unbounded scenes. Furthermore, with our resampling technique for grid-based methods, we can increase the number of valid samples to train NeRF volume. We extensively evaluate our method in our newly introduced synthetic and real-world egocentric 360 video datasets, and it consistently achieves state-of-the-art performance.
****

Box-Level Active Detection

Mengyao Lyu, Jundong Zhou, Hui Chen, Yijie Huang, Dongdong Yu, Yaqian Li, Yandong Guo, Yuchen Guo, Liuyu Xiang, Guiguang Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23766-23775

Active learning selects informative samples for annotation within budget, which has proven efficient recently on object detection. However, the widely used active detection benchmarks conduct image-level evaluation, which is unrealistic in human workload estimation and biased towards crowded images. Furthermore, existing methods still perform image-level annotation, but equally scoring all targets within the same image incurs waste of budget and redundant labels. Having revealed above problems and limitations, we introduce a box-level active detection framework that controls a box-based budget per cycle, prioritizes informative targets and avoids redundancy for fair comparison and efficient application. Under the proposed box-level setting, we devise a novel pipeline, namely Complementary Pseudo Active Strategy (ComPAS). It exploits both human annotations and the model intelligence in a complementary fashion: an efficient input-end committee queries labels for informative objects only; meantime well-learned targets are identified by the model and compensated with pseudo-labels. ComPAS consistently outperforms 10 competitors under 4 settings in a unified codebase. With supervision from labeled data only, it achieves 100% supervised performance of VOC0712 with merely 19% box annotations. On the COCO dataset, it yields up to 4.3% mAP improvement over the second-best method. ComPAS also supports training with the unlabeled pool, where it surpasses 90% COCO supervised performance with 85% label reduction. Our source code is publicly available at https://github.com/lyumengyao/blad.
****

Coreset Sampling From Open-Set for Fine-Grained Self-Supervised Learning

Sungnyun Kim, Sangmin Bae, Se-Young Yun; Proceedings of the IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7537-7547
Deep learning in general domains has constantly been extended to domain-specific tasks requiring the recognition of fine-grained characteristics. However, real-world applications for fine-grained tasks suffer from two challenges: a high reliance on expert knowledge for annotation and necessity of a versatile model for various downstream tasks in a specific domain (e.g., prediction of categories, bounding boxes, or pixel-wise annotations). Fortunately, the recent self-supervised learning (SSL) is a promising approach to pretrain a model without annotations, serving as an effective initialization for any downstream tasks. Since SSL does not rely on the presence of annotation, in general, it utilizes the large-scale unlabeled dataset, referred to as an open-set. In this sense, we introduce a novel Open-Set Self-Supervised Learning problem under the assumption that a large-scale unlabeled open-set is available, as well as the fine-grained target dataset, during a pretraining phase. In our problem setup, it is crucial to consider the distribution mismatch between the open-set and target dataset. Hence, we propose SimCore algorithm to sample a coreset, the subset of an open-set that has a minimum distance to the target dataset in the latent space. We demonstrate that SimCore significantly improves representation learning performance through extensive experimental settings, including eleven fine-grained datasets and seven open-sets in various downstream tasks.
****
Trace and Pace: Controllable Pedestrian Animation via Guided Trajectory Diffusion
Davis Rempe, Zhengyi Luo, Xue Bin Peng, Ye Yuan, Kris Kitani, Karsten Kreis, Sanja Fidler, Or Litany; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13756-13766
We introduce a method for generating realistic pedestrian trajectories and full-body animations that can be controlled to meet user-defined goals. We draw on recent advances in guided diffusion modeling to achieve test-time controllability of trajectories, which is normally only associated with rule-based systems. Our guided diffusion model allows users to constrain trajectories through target waypoints, speed, and specified social groups while accounting for the surrounding environment context. This trajectory diffusion model is integrated with a novel physics-based humanoid controller to form a closed-loop, full-body pedestrian animation system capable of placing large crowds in a simulated environment with varying terrains. We further propose utilizing the value function learned during RL training of the animation controller to guide diffusion to produce trajectories better suited for particular scenarios such as collision avoidance and traversing uneven terrain.
****
Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability
Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, Olga Russakovsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10932-10941
Concept-based interpretability methods aim to explain a deep neural network model's components and predictions using a pre-defined set of semantic concepts. These methods evaluate a trained model on a new, "probe" dataset and correlate the model's outputs with concepts labeled in that dataset. Despite their popularity, they suffer from limitations that are not well-understood and articulated in the literature. In this work, we identify and analyze three commonly overlooked factors in concept-based explanations. First, we find that the choice of the probe dataset has a profound impact on the generated explanations. Our analysis reveals that different probe datasets lead to very different explanations, suggesting that the generated explanations are not generalizable outside the probe dataset. Second, we find that concepts in the probe dataset are often harder to learn than the target classes they are used to explain, calling into question the correctness of the explanations. We argue that only easily learnable concepts should be used in concept-based explanations. Finally, while existing methods use hundreds or even thousands of concepts, our human studies reveal a much stricter uppe

r bound of 32 concepts or less, beyond which the explanations are much less prac tically useful. We discuss the implications of our findings and provide suggesti ons for future development of concept-based interpretability methods. Code for o ur analysis and user interface can be found at https://github.com/princetonvisua lai/OverlookedFactors.
****
Unsupervised 3D Shape Reconstruction by Part Retrieval and Assembly
Xianghao Xu, Paul Guerrero, Matthew Fisher, Siddhartha Chaudhuri, Daniel Ritchie ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2023, pp. 8559-8567
Representing a 3D shape with a set of primitives can aid perception of structure , improve robotic object manipulation, and enable editing, stylization, and comp ression of 3D shapes. Existing methods either use simple parametric primitives o r learn a generative shape space of parts. Both have limitations: parametric pri mitives lead to coarse approximations, while learned parts offer too little cont rol over the decomposition. We instead propose to decompose shapes using a libra ry of 3D parts provided by the user, giving full control over the choice of part s. The library can contain parts with high-quality geometry that are suitable fo r a given category, resulting in meaningful decom- positions with clean geometry . The type of decomposition can also be controlled through the choice of parts i n the library. Our method works via a unsupervised approach that iteratively ret rieves parts from the library and refines their placements. We show that this ap proach gives higher reconstruction accuracy and more desirable decompositions th an existing approaches. Additionally, we show how the decom- position can be con trolled through the part library by using different part libraries to reconstruc t the same shapes.
****
SeqTrack: Sequence to Sequence Learning for Visual Object Tracking
Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, Han Hu; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14572- 14581
In this paper, we present a new sequence-to-sequence learning framework for visu al tracking, dubbed SeqTrack. It casts visual tracking as a sequence generation problem, which predicts object bounding boxes in an autoregressive fashion. This is different from prior Siamese trackers and transformer trackers, which rely o n designing complicated head networks, such as classification and regression hea ds. SeqTrack only adopts a simple encoder-decoder transformer architecture. The encoder extracts visual features with a bidirectional transformer, while the dec oder generates a sequence of bounding box values autoregressively with a causal transformer. The loss function is a plain cross-entropy. Such a sequence learnin g paradigm not only simplifies tracking framework, but also achieves competitive performance on benchmarks. For instance, SeqTrack gets 72.5% AUC on LaSOT, esta blishing a new state-of-the-art performance. Code and models are available at ht tps://github.com/microsoft/VideoX.
****
Self-Supervised Non-Uniform Kernel Estimation With Flow-Based Motion Prior for B lind Image Deblurring
Zhenxuan Fang, Fangfang Wu, Weisheng Dong, Xin Li, Jinjian Wu, Guangming Shi; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18105-18114
Many deep learning-based solutions to blind image deblurring estimate the blur r epresentation and reconstruct the target image from its blurry observation. Howe ver, these methods suffer from severe performance degradation in real-world scen arios because they ignore important prior information about motion blur (e.g., r eal-world motion blur is diverse and spatially varying). Some methods have attem pted to explicitly estimate non-uniform blur kernels by CNNs, but accurate estim ation is still challenging due to the lack of ground truth about spatially varyi ng blur kernels in real-world images. To address these issues, we propose to rep resent the field of motion blur kernels in a latent space by normalizing flows, and design CNNs to predict the latent codes instead of motion kernels. To furthe

r improve the accuracy and robustness of non-uniform kernel estimation, we introduce uncertainty learning into the process of estimating latent codes and propose a multi-scale kernel attention module to better integrate image features with estimated kernels. Extensive experimental results, especially on real-world blur datasets, demonstrate that our method achieves state-of-the-art results in terms of both subjective and objective quality as well as excellent generalization performance for non-uniform image deblurring. The code is available at https://see.xidian.edu.cn/faculty/wsdong/Projects/UFPNet.htm.
****
AutoLabel: CLIP-Based Framework for Open-Set Video Domain Adaptation
Giacomo Zara, Subhankar Roy, Paolo Rota, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11504-11513
Open-set Unsupervised Video Domain Adaptation (OUVDA) deals with the task of adapting an action recognition model from a labelled source domain to an unlabelled target domain that contains "target-private" categories, which are present in the target but absent in the source. In this work we deviate from the prior work of training a specialized open-set classifier or weighted adversarial learning by proposing to use pre-trained Language and Vision Models (CLIP). The CLIP is well suited for OUVDA due to its rich representation and the zero-shot recognition capabilities. However, rejecting target-private instances with the CLIP's zero-shot protocol requires oracle knowledge about the target-private label names. To circumvent the impossibility of the knowledge of label names, we propose AutoLabel that automatically discovers and generates object-centric compositional candidate target-private class names. Despite its simplicity, we show that CLIP when equipped with AutoLabel can satisfactorily reject the target-private instances, thereby facilitating better alignment between the shared classes of the two domains. The code is available.
****
Generative Semantic Segmentation
Jiaqi Chen, Jiachen Lu, Xiatian Zhu, Li Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7111-7120
We present Generative Semantic Segmentation (GSS), a generative learning approach for semantic segmentation. Uniquely, we cast semantic segmentation as an image-conditioned mask generation problem. This is achieved by replacing the conventional per-pixel discriminative learning with a latent prior learning process. Specifically, we model the variational posterior distribution of latent variables given the segmentation mask. To that end, the segmentation mask is expressed with a special type of image (dubbed as maskige). This posterior distribution allows to generate segmentation masks unconditionally. To achieve semantic segmentation on a given image, we further introduce a conditioning network. It is optimized by minimizing the divergence between the posterior distribution of maskige (i.e., segmentation masks) and the latent prior distribution of input training images. Extensive experiments on standard benchmarks show that our GSS can perform competitively to prior art alternatives in the standard semantic segmentation setting, whilst achieving a new state of the art in the more challenging cross-domain setting.
****
Instant-NVR: Instant Neural Volumetric Rendering for Human-Object Interactions From Monocular RGBD Stream
Yuheng Jiang, Kaixin Yao, Zhuo Su, Zhehao Shen, Haimin Luo, Lan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 595-605
Convenient 4D modeling of human-object interactions is essential for numerous applications. However, monocular tracking and rendering of complex interaction scenarios remain challenging. In this paper, we propose Instant-NVR, a neural approach for instant volumetric human-object tracking and rendering using a single RGBD camera. It bridges traditional non-rigid tracking with recent instant radiance field techniques via a multi-thread tracking-rendering mechanism. In the tracking front-end, we adopt a robust human-object capture scheme to provide sufficie

nt motion priors. We further introduce a separated instant neural representation with a novel hybrid deformation module for the interacting scene. We also provide an on-the-fly reconstruction scheme of the dynamic/static radiance fields via efficient motion-prior searching. Moreover, we introduce an online key frame selection scheme and a rendering-aware refinement strategy to significantly improve the appearance details for online novel-view synthesis. Extensive experiments demonstrate the effectiveness and efficiency of our approach for the instant generation of human-object radiance fields on the fly, notably achieving real-time photo-realistic novel view synthesis under complex human-object interactions.
****
Aligning Step-by-Step Instructional Diagrams to Video Demonstrations
Jiahao Zhang, Anoop Cherian, Yanbin Liu, Yizhak Ben-Shabat, Cristian Rodriguez, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2483-2492
Multimodal alignment facilitates the retrieval of instances from one modality when queried using another. In this paper, we consider a novel setting where such an alignment is between (i) instruction steps that are depicted as assembly diagrams (commonly seen in Ikea assembly manuals) and (ii) video segments from in-the-wild videos; these videos comprising an enactment of the assembly actions in the real world. To learn this alignment, we introduce a novel supervised contrastive learning method that learns to align videos with the subtle details in the assembly diagrams, guided by a set of novel losses. To study this problem and demonstrate the effectiveness of our method, we introduce a novel dataset: IAW---for Ikea assembly in the wild---consisting of 183 hours of videos from diverse furniture assembly collections and nearly 8,300 illustrations from their associated instruction manuals and annotated for their ground truth alignments. We define two tasks on this dataset: First, nearest neighbor retrieval between video segments and illustrations, and, second, alignment of instruction steps and the segments for each video. Extensive experiments on IAW demonstrate superior performances of our approach against alternatives.
****
Collecting Cross-Modal Presence-Absence Evidence for Weakly-Supervised Audio-Visual Event Perception
Junyu Gao, Mengyuan Chen, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18827-18836
With only video-level event labels, this paper targets at the task of weakly-supervised audio-visual event perception (WS-AVEP), which aims to temporally localize and categorize events belonging to each modality. Despite the recent progress, most existing approaches either ignore the unsynchronized property of audio-visual tracks or discount the complementary modality for explicit enhancement. We argue that, for an event residing in one modality, the modality itself should provide ample presence evidence of this event, while the other complementary modality is encouraged to afford the absence evidence as a reference signal. To this end, we propose to collect Cross-Modal Presence-Absence Evidence (CMPAE) in a unified framework. Specifically, by leveraging uni-modal and cross-modal representations, a presence-absence evidence collector (PAEC) is designed under Subjective Logic theory. To learn the evidence in a reliable range, we propose a joint-modal mutual learning (JML) process, which calibrates the evidence of diverse audible, visible, and audi-visible events adaptively and dynamically. Extensive experiments show that our method surpasses state-of-the-arts (e.g., absolute gains of 3.6% and 6.1% in terms of event-level visual and audio metrics). Code is available in github.com/MengyuanChen21/CVPR2023-CMPAE.
****
High-Fidelity and Freely Controllable Talking Head Video Generation
Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5609-5619
Talking head generation is to generate video based on a given source identity and target motion. However, current methods face several challenges that limit the quality and controllability of the generated videos. First, the generated face

often has unexpected deformation and severe distortions. Second, the driving ima
ge does not explicitly disentangle movement-relevant information, such as poses
and expressions, which restricts the manipulation of different attributes during
 generation. Third, the generated videos tend to have flickering artifacts due t
o the inconsistency of the extracted landmarks between adjacent frames. In this
paper, we propose a novel model that produces high-fidelity talking head videos
with free control over head pose and expression. Our method leverages both self-
supervised learned landmarks and 3D face model-based landmarks to model the moti
on. We also introduce a novel motion-aware multi-scale feature alignment module
to effectively transfer the motion without face distortion. Furthermore, we enha
nce the smoothness of the synthesized talking head videos with a feature context
 adaptation and propagation module. We evaluate our model on challenging dataset
s and demonstrate its state-of-the-art performance. More information is availabl
e at https://yuegao.me/PECHead.
****
Q-DETR: An Efficient Low-Bit Quantized Detection Transformer
Sheng Xu, Yanjing Li, Mingbao Lin, Peng Gao, Guodong Guo, Jinhu Lü, Baochang Zha
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2023, pp. 3842-3851
The recent detection transformer (DETR) has advanced object detection, but its a
pplication on resource-constrained devices requires massive computation and memo
ry resources. Quantization stands out as a solution by representing the network
in low-bit parameters and operations. However, there is a significant performanc
e drop when performing low-bit quantized DETR (Q-DETR) with existing quantizatio
n methods. We find that the bottlenecks of Q-DETR come from the query informatio
n distortion through our empirical analyses. This paper addresses this problem b
ased on a distribution rectification distillation (DRD). We formulate our DRD as
 a bi-level optimization problem, which can be derived by generalizing the infor
mation bottleneck (IB) principle to the learning of Q-DETR. At the inner level,
we conduct a distribution alignment for the queries to maximize the self-informa
tion entropy. At the upper level, we introduce a new foreground-aware query matc
hing scheme to effectively transfer the teacher information to distillation-desi
red features to minimize the conditional information entropy. Extensive experime
ntal results show that our method performs much better than prior arts. For exam
ple, the 4-bit Q-DETR can theoretically accelerate DETR with ResNet-50 backbone
by 6.6x and achieve 39.4% AP, with only 2.6% performance gaps than its real-valu
ed counterpart on the COCO dataset.
****
DINER: Depth-Aware Image-Based NEural Radiance Fields
Malte Prinzler, Otmar Hilliges, Justus Thies; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12449-12459
We present Depth-aware Image-based NEural Radiance fields (DINER). Given a spars
e set of RGB input views, we predict depth and feature maps to guide the reconst
ruction of a volumetric scene representation that allows us to render 3D objects
 under novel views. Specifically, we propose novel techniques to incorporate dep
th information into feature fusion and efficient scene sampling. In comparison t
o the previous state of the art, DINER achieves higher synthesis quality and can
 process input views with greater disparity. This allows us to capture scenes mo
re completely without changing capturing hardware requirements and ultimately en
ables larger viewpoint changes during novel view synthesis. We evaluate our meth
od by synthesizing novel views, both for human heads and for general objects, an
d observe significantly improved qualitative results and increased perceptual me
trics compared to the previous state of the art.
****
Burstormer: Burst Image Restoration and Enhancement Transformer
Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, Ming-Hsuan Ya
ng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2023, pp. 5703-5712
On a shutter press, modern handheld cameras capture multiple images in rapid suc
cession and merge them to generate a single image. However, individual frames in

a burst are misaligned due to inevitable motions and contain multiple degradations. The challenge is to properly align the successive image shots and merge their complimentary information to achieve high-quality outputs. Towards this direction, we propose Burstormer: a novel transformer-based architecture for burst image restoration and enhancement. In comparison to existing works, our approach exploits multi-scale local and non-local features to achieve improved alignment and feature fusion. Our key idea is to enable inter-frame communication in the burst neighborhoods for information aggregation and progressive fusion while modeling the burst-wide context. However, the input burst frames need to be properly aligned before fusing their information. Therefore, we propose an enhanced deformable alignment module for aligning burst features with regards to the reference frame. Unlike existing methods, the proposed alignment module not only aligns burst features but also exchanges feature information and maintains focused communication with the reference frame through the proposed reference-based feature enrichment mechanism, which facilitates handling complex motions. After multi-level alignment and enrichment, we re-emphasize on inter-frame communication within burst using a cyclic burst sampling module. Finally, the inter-frame information is aggregated using the proposed burst feature fusion module followed by progressive upsampling. Our Burstormer outperforms state-of-the-art methods on burst super-resolution, burst denoising and burst low-light enhancement. Our codes and pre-trained models are available at https://github.com/akshaydudhane16/Burstormer.
****
Progressive Transformation Learning for Leveraging Virtual Images in Training
Yi-Ting Shen, Hyungtae Lee, Heesung Kwon, Shuvra S. Bhattacharyya; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 835-844
To effectively interrogate UAV-based images for detecting objects of interest, such as humans, it is essential to acquire large-scale UAV-based datasets that include human instances with various poses captured from widely varying viewing angles. As a viable alternative to laborious and costly data curation, we introduce Progressive Transformation Learning (PTL), which gradually augments a training dataset by adding transformed virtual images with enhanced realism. Generally, a virtual2real transformation generator in the conditional GAN framework suffers from quality degradation when a large domain gap exists between real and virtual images. To deal with the domain gap, PTL takes a novel approach that progressively iterates the following three steps: 1) select a subset from a pool of virtual images according to the domain gap, 2) transform the selected virtual images to enhance realism, and 3) add the transformed virtual images to the training set while removing them from the pool. In PTL, accurately quantifying the domain gap is critical. To do that, we theoretically demonstrate that the feature representation space of a given object detector can be modeled as a multivariate Gaussian distribution from which the Mahalanobis distance between a virtual object and the Gaussian distribution of each object category in the representation space can be readily computed. Experiments show that PTL results in a substantial performance increase over the baseline, especially in the small data and the cross-domain regime.
****
Co-Speech Gesture Synthesis by Reinforcement Learning With Contrastive Pre-Trained Rewards
Mingyang Sun, Mengchen Zhao, Yaqing Hou, Minglei Li, Huang Xu, Songcen Xu, Jianye Hao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2331-2340
There is a growing demand of automatically synthesizing co-speech gestures for virtual characters. However, it remains a challenge due to the complex relationship between input speeches and target gestures. Most existing works focus on predicting the next gesture that fits the data best, however, such methods are myopic and lack the ability to plan for future gestures. In this paper, we propose a novel reinforcement learning (RL) framework called RACER to generate sequences of gestures that maximize the overall satisfactory. RACER employs a vector quanti

zed variational autoencoder to learn compact representations of gestures and a G PT-based policy architecture to generate coherent sequence of gestures autoregre ssively. In particular, we propose a contrastive pre-training approach to calcul ate the rewards, which integrates contextual information into action evaluation and successfully captures the complex relationships between multi-modal speech-g esture data. Experimental results show that our method significantly outperforms existing baselines in terms of both objective metrics and subjective human judg ements. Demos can be found at https://github.com/RLracer/RACER.git.
****
Reconstructing Signing Avatars From Video Using Linguistic Priors
Maria-Paola Forte, Peter Kulits, Chun-Hao P. Huang, Vasileios Choutas, Dimitrios Tzionas, Katherine J. Kuchenbecker, Michael J. Black; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12791 -12801
Sign language (SL) is the primary method of communication for the 70 million Dea f people around the world. Video dictionaries of isolated signs are a core SL le arning tool. Replacing these with 3D avatars can aid learning and enable AR/VR a pplications, improving access to technology and online media. However, little wo rk has attempted to estimate expressive 3D avatars from SL video; occlusion, noi se, and motion blur make this task difficult. We address this by introducing nov el linguistic priors that are universally applicable to SL and provide constrain ts on 3D hand pose that help resolve ambiguities within isolated signs. Our meth od, SGNify, captures fine-grained hand pose, facial expression, and body movemen t fully automatically from in-the-wild monocular SL videos. We evaluate SGNify q uantitatively by using a commercial motion-capture system to compute 3D avatars synchronized with monocular video. SGNify outperforms state-of-the-art 3D body-p ose- and shape-estimation methods on SL videos. A perceptual study shows that SG Nify's 3D reconstructions are significantly more comprehensible and natural than those of previous methods and are on par with the source videos. Code and data are available at sgnify.is.tue.mpg.de.
****
DeepMapping2: Self-Supervised Large-Scale LiDAR Map Optimization
Chao Chen, Xinhao Liu, Yiming Li, Li Ding, Chen Feng; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9306-9 316
LiDAR mapping is important yet challenging in self-driving and mobile robotics. To tackle such a global point cloud registration problem, DeepMapping converts t he complex map estimation into a self-supervised training of simple deep network s. Despite its broad convergence range on small datasets, DeepMapping still cann ot produce satisfactory results on large-scale datasets with thousands of frames . This is due to the lack of loop closures and exact cross-frame point correspon dences, and the slow convergence of its global localization network. We propose DeepMapping2 by adding two novel techniques to address these issues: (1) organiz ation of training batch based on map topology from loop closing, and (2) self-su pervised local-to-global point consistency loss leveraging pairwise registration . Our experiments and ablation studies on public datasets such as KITTI, NCLT, a nd Nebula, demonstrate the effectiveness of our method.
****
SDC-UDA: Volumetric Unsupervised Domain Adaptation Framework for Slice-Direction Continuous Cross-Modality Medical Image Segmentation
Hyungseob Shin, Hyeongyu Kim, Sewon Kim, Yohan Jun, Taejoon Eo, Dosik Hwang; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7412-7421
Recent advances in deep learning-based medical image segmentation studies achiev e nearly human-level performance in fully supervised manner. However, acquiring pixel-level expert annotations is extremely expensive and laborious in medical i maging fields. Unsupervised domain adaptation (UDA) can alleviate this problem, which makes it possible to use annotated data in one imaging modality to train a network that can successfully perform segmentation on target imaging modality w ith no labels. In this work, we propose SDC-UDA, a simple yet effective volumetr

ic UDA framework for Slice-Direction Continuous cross-modality medical image segmentation which combines intra- and inter-slice self-attentive image translation, uncertainty-constrained pseudo-label refinement, and volumetric self-training. Our method is distinguished from previous methods on UDA for medical image segmentation in that it can obtain continuous segmentation in the slice direction, thereby ensuring higher accuracy and potential in clinical practice. We validate SDC-UDA with multiple publicly available cross-modality medical image segmentation datasets and achieve state-of-the-art segmentation performance, not to mention the superior slice-direction continuity of prediction compared to previous studies.
****
DoNet: Deep De-Overlapping Network for Cytology Instance Segmentation
Hao Jiang, Rushan Zhang, Yanning Zhou, Yumeng Wang, Hao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15641-15650
Cell instance segmentation in cytology images has significant importance for biology analysis and cancer screening, while remains challenging due to 1) the extensive overlapping translucent cell clusters that cause the ambiguous boundaries, and 2) the confusion of mimics and debris as nuclei. In this work, we proposed a De-overlapping Network (DoNet) in a decompose-and-recombined strategy. A Dual-path Region Segmentation Module (DRM) explicitly decomposes the cell clusters into intersection and complement regions, followed by a Semantic Consistency-guided Recombination Module (CRM) for integration. To further introduce the containment relationship of the nucleus in the cytoplasm, we design a Mask-guided Region Proposal Strategy (MRP) that integrates the cell attention maps for inner-cell instance prediction. We validate the proposed approach on ISBI2014 and CPS datasets. Experiments show that our proposed DoNet significantly outperforms other state-of-the-art (SOTA) cell instance segmentation methods. The code is available at https://github.com/DeepDoNet/DoNet.
****
AVFace: Towards Detailed Audio-Visual 4D Face Reconstruction
Aggelina Chatziagapi, Dimitris Samaras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16878-16889
In this work, we present a multimodal solution to the problem of 4D face reconstruction from monocular videos. 3D face reconstruction from 2D images is an under-constrained problem due to the ambiguity of depth. State-of-the-art methods try to solve this problem by leveraging visual information from a single image or video, whereas 3D mesh animation approaches rely more on audio. However, in most cases (e.g. AR/VR applications), videos include both visual and speech information. We propose AVFace that incorporates both modalities and accurately reconstructs the 4D facial and lip motion of any speaker, without requiring any 3D ground truth for training. A coarse stage estimates the per-frame parameters of a 3D morphable model, followed by a lip refinement, and then a fine stage recovers facial geometric details. Due to the temporal audio and video information captured by transformer-based modules, our method is robust in cases when either modality is insufficient (e.g. face occlusions). Extensive qualitative and quantitative evaluation demonstrates the superiority of our method over the current state-of-the-art.
****
Divide and Conquer: Answering Questions With Object Factorization and Compositional Reasoning
Shi Chen, Qi Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6736-6745
Humans have the innate capability to answer diverse questions, which is rooted in the natural ability to correlate different concepts based on their semantic relationships and decompose difficult problems into sub-tasks. On the contrary, existing visual reasoning methods assume training samples that capture every possible object and reasoning problem, and rely on black-boxed models that commonly exploit statistical priors. They have yet to develop the capability to address novel objects or spurious biases in real-world scenarios, and also fall short of i

nterpreting the rationales behind their decisions. Inspired by humans' reasoning of the visual world, we tackle the aforementioned challenges from a compositional perspective, and propose an integral framework consisting of a principled object factorization method and a novel neural module network. Our factorization method decomposes objects based on their key characteristics, and automatically derives prototypes that represent a wide range of objects. With these prototypes encoding important semantics, the proposed network then correlates objects by measuring their similarity on a common semantic space and makes decisions with a compositional reasoning process. It is capable of answering questions with diverse objects regardless of their availability during training, and overcoming the issues of biased question-answer distributions. In addition to the enhanced generalizability, our framework also provides an interpretable interface for understanding the decision-making process of models. Our code is available at https://github.com/szzexpoi/POEM.
****
Instant Domain Augmentation for LiDAR Semantic Segmentation
Kwonyoung Ryu, Soonmin Hwang, Jaesik Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9350-9360
Despite the increasing popularity of LiDAR sensors, perception algorithms using 3D LiDAR data struggle with the 'sensor-bias problem'. Specifically, the performance of perception algorithms significantly drops when an unseen specification of LiDAR sensor is applied at test time due to the domain discrepancy. This paper presents a fast and flexible LiDAR augmentation method for the semantic segmentation task, called 'LiDomAug'. It aggregates raw LiDAR scans and creates a LiDAR scan of any configurations with the consideration of dynamic distortion and occlusion, resulting in instant domain augmentation. Our on-demand augmentation module runs at 330 FPS, so it can be seamlessly integrated into the data loader in the learning framework. In our experiments, learning-based approaches aided with the proposed LiDomAug are less affected by the sensor-bias issue and achieve new state-of-the-art domain adaptation performances on SemanticKITTI and nuScenes dataset without the use of the target domain data. We also present a sensor-agnostic model that faithfully works on the various LiDAR configurations.
****
A Characteristic Function-Based Method for Bottom-Up Human Pose Estimation
Haoxuan Qu, Yujun Cai, Lin Geng Foo, Ajay Kumar, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13009-13018
Most recent methods formulate the task of human pose estimation as a heatmap estimation problem, and use the overall L2 loss computed from the entire heatmap to optimize the heatmap prediction. In this paper, we show that in bottom-up human pose estimation where each heatmap often contains multiple body joints, using the overall L2 loss to optimize the heatmap prediction may not be the optimal choice. This is because, minimizing the overall L2 loss cannot always lead the model to locate all the body joints across different sub-regions of the heatmap more accurately. To cope with this problem, from a novel perspective, we propose a new bottom-up human pose estimation method that optimizes the heatmap prediction via minimizing the distance between two characteristic functions respectively constructed from the predicted heatmap and the groundtruth heatmap. Our analysis presented in this paper indicates that the distance between these two characteristic functions is essentially the upper bound of the L2 losses w.r.t. sub-regions of the predicted heatmap. Therefore, via minimizing the distance between the two characteristic functions, we can optimize the model to provide a more accurate localization result for the body joints in different sub-regions of the predicted heatmap. We show the effectiveness of our proposed method through extensive experiments on the COCO dataset and the CrowdPose dataset.
****
SceneTrilogy: On Human Scene-Sketch and Its Complementarity With Photo and Text
Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10972-10983

In this paper, we extend scene understanding to include that of human sketch. The result is a complete trilogy of scene representation from three diverse and complementary modalities -- sketch, photo, and text. Instead of learning a rigid three-way embedding and be done with it, we focus on learning a flexible joint embedding that fully supports the "optionality" that this complementarity brings. Our embedding supports optionality on two axis: (i) optionality across modalities -- use any combination of modalities as query for downstream tasks like retrieval, (ii) optionality across tasks -- simultaneously utilising the embedding for either discriminative (e.g., retrieval) or generative tasks (e.g., captioning). This provides flexibility to end-users by exploiting the best of each modality, therefore serving the very purpose behind our proposal of a trilogy at the first place. First, a combination of information-bottleneck and conditional invertible neural networks disentangle the modality-specific component from modality-agnostic in sketch, photo, and text. Second, the modality-agnostic instances from sketch, photo, and text are synergised using a modified cross-attention. Once learned, we show our embedding can accommodate a multi-facet of scene-related tasks, including those enabled for the first time by the inclusion of sketch, all without any task-specific modifications. Project Page: http://www.pinakinathc.me/scenetrilogy
****
ERM-KTP: Knowledge-Level Machine Unlearning via Knowledge Transfer
Shen Lin, Xiaoyu Zhang, Chenyang Chen, Xiaofeng Chen, Willy Susilo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20147-20155
Machine unlearning can fortify the privacy and security of machine learning applications. Unfortunately, the exact unlearning approaches are inefficient, and the approximate unlearning approaches are unsuitable for complicated CNNs. Moreover, the approximate approaches have serious security flaws because even unlearning completely different data points can produce the same contribution estimation as unlearning the target data points. To address the above problems, we try to define machine unlearning from the knowledge perspective, and we propose a knowledge-level machine unlearning method, namely ERM-KTP. Specifically, we propose an entanglement-reduced mask (ERM) structure to reduce the knowledge entanglement among classes during the training phase. When receiving the unlearning requests, we transfer the knowledge of the non-target data points from the original model to the unlearned model and meanwhile prohibit the knowledge of the target data points via our proposed knowledge transfer and prohibition (KTP) method. Finally, we will get the unlearned model as the result and delete the original model to accomplish the unlearning process. Especially, our proposed ERM-KTP is an interpretable unlearning method because the ERM structure and the crafted masks in KTP can explicitly explain the operation and the effect of unlearning data points. Extensive experiments demonstrate the effectiveness, efficiency, high fidelity, and scalability of the ERM-KTP unlearning method.
****
RefSR-NeRF: Towards High Fidelity and Super Resolution View Synthesis
Xudong Huang, Wei Li, Jie Hu, Hanting Chen, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8244-8253
We present Reference-guided Super-Resolution Neural Radiance Field (RefSR-NeRF) that extends NeRF to super resolution and photorealistic novel view synthesis. Despite NeRF's extraordinary success in the neural rendering field, it suffers from blur in high resolution rendering because its inherent multilayer perceptron struggles to learn high frequency details and incurs a computational explosion as resolution increases. Therefore, we propose RefSR-NeRF, an end-to-end framework that first learns a low resolution NeRF representation, and then reconstructs the high frequency details with the help of a high resolution reference image. We observe that simply introducing the pre-trained models from the literature tends to produce unsatisfied artifacts due to the divergence in the degradation model. To this end, we design a novel lightweight RefSR model to learn the inverse degradation process from NeRF renderings to target HR ones. Extensive experiment

s on multiple benchmarks demonstrate that our method exhibits an impressive trade-off among rendering quality, speed, and memory usage, outperforming or on par with NeRF and its variants while being 52x speedup with minor extra memory usage.
****
DATE: Domain Adaptive Product Seeker for E-Commerce
Haoyuan Li, Hao Jiang, Tao Jin, Mengyan Li, Yan Chen, Zhijie Lin, Yang Zhao, Zhou Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19315-19324
Product Retrieval (PR) and Grounding (PG), aiming to seek image and object-level products respectively according to a textual query, have attracted great interest recently for better shopping experience. Owing to the lack of relevant datasets, we collect two large-scale benchmark datasets from Taobao Mall and Live domains with about 474k and 101k image-query pairs for PR, and manually annotate the object bounding boxes in each image for PG. As annotating boxes is expensive and time-consuming, we attempt to transfer knowledge from annotated domain to unannotated for PG to achieve un-supervised Domain Adaptation (PG-DA). We propose a Domain Adaptive producT sEeker (DATE) framework, regarding PR and PG as Product Seeking problem at different levels, to assist the query date the product. Concretely, we first design a semantics-aggregated feature extractor for each modality to obtain concentrated and comprehensive features for following efficient retrieval and fine-grained grounding tasks. Then, we present two cooperative seekers to simultaneously search the image for PR and localize the product for PG. Besides, we devise a domain aligner for PG-DA to alleviate uni-modal marginal and multi-modal conditional distribution shift between source and target domains, and design a pseudo box generator to dynamically select reliable instances and generate bounding boxes for further knowledge transfer. Extensive experiments show that our DATE achieves satisfactory performance in fully-supervised PR, PG and un-supervised PG-DA. Our desensitized datasets will be publicly available here https://github.com/Taobao-live/Product-Seeking.
****
Polarimetric iToF: Measuring High-Fidelity Depth Through Scattering Media
Daniel S. Jeon, Andréas Meuleman, Seung-Hwan Baek, Min H. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12353-12362
Indirect time-of-flight (iToF) imaging allows us to capture dense depth information at a low cost. However, iToF imaging often suffers from multipath interference (MPI) artifacts in the presence of scattering media, resulting in severe depth-accuracy degradation. For instance, iToF cameras cannot measure depth accurately through fog because ToF active illumination scatters back to the sensor before reaching the farther target surface. In this work, we propose a polarimetric iToF imaging method that can capture depth information robustly through scattering media. Our observations on the principle of indirect ToF imaging and polarization of light allow us to formulate a novel computational model of scattering-aware polarimetric phase measurements that enables us to correct MPI errors. We first devise a scattering-aware polarimetric iToF model that can estimate the phase of unpolarized backscattered light. We then combine the optical filtering of polarization and our computational modeling of unpolarized backscattered light via scattering analysis of phase and amplitude. This allows us to tackle the MPI problem by estimating the scattering energy through the participating media. We validate our method on an experimental setup using a customized off-the-shelf iToF camera. Our method outperforms baseline methods by a significant margin by means of our scattering model and polarimetric phase measurements.
****
Jedi: Entropy-Based Localization and Removal of Adversarial Patches
Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, Ihsen Alouani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4087-4095
Real-world adversarial physical patches were recently shown to be successful in compromising state-of-the-art models in a variety of computer vision application

s. The most promising defenses that are based on either input gradient or featur es analyses have been shown to be compromised by recent GAN-based adaptive attac ks that generate realistic/naturalistic patches. In this paper, we propose Jedi, a new defense against adversarial patches that is resilient to realistic patch attacks, and also improves detection and recovery compared to the state of the a rt. Jedi leverages two new ideas: (1) it improves the identification of potentia l patch regions using entropy analysis: we show that the entropy of adversarial patches is high, even in naturalistic patches; and (2) it improves the localizat ion of adversarial patches, using an autoencoder that is able to complete patch regions and filter out normal regions with high entropy that are not part of a p atch. Jedi achieves high precision adversarial patch localization, which we show is critical to successfully repair the images. Since Jedi relies on an input en tropy analysis, it is model-agnostic, and can be applied on pre-trained off-the- shelf models without changes to the training or inference of the protected model s. Jedi detects on average 90% of adversarial patches across different benchmark s and recovers up to 94% of successful patch attacks (Compared to 75% and 65% fo r LGS and Jujutsu, respectively). Jedi is also able to continue detection even i n the presence of adaptive realistic patches that are able to fool other defense s.
****
Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomo us Driving
Abdul Hannan Khan, Mohammed Shariq Nawaz, Andreas Dengel; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 54 76-5485
Autonomous driving systems rely heavily on the underlying perception module whic h needs to be both performant and efficient to allow precise decisions in real-t ime. Avoiding collisions with pedestrians is of topmost priority in any autonomo us driving system. Therefore, pedestrian detection is one of the core parts of s uch systems' perception modules. Current state-of-the-art pedestrian detectors h ave two major issues. Firstly, they have long inference times which affect the e fficiency of the whole perception module, and secondly, their performance in the case of small and heavily occluded pedestrians is poor. We propose Localized Se mantic Feature Mixers (LSFM), a novel, anchor-free pedestrian detection architec ture. It uses our novel Super Pixel Pyramid Pooling module instead of the, compu tationally costly, Feature Pyramid Networks for feature encoding. Moreover, our MLPMixer-based Dense Focal Detection Network is used as a light detection head, reducing computational effort and inference time compared to existing approaches . To boost the performance of the proposed architecture, we adapt and use mixup augmentation which improves the performance, especially in small and heavily occ luded cases. We benchmark LSFM against the state-of-the-art on well-established traffic scene pedestrian datasets. The proposed LSFM achieves state-of-the-art p erformance in Caltech, City Persons, Euro City Persons, and TJU-Traffic-Pedestri an datasets while reducing the inference time on average by 55%. Further, LSFM b eats the human baseline for the first time in the history of pedestrian detectio n. Finally, we conducted a cross-dataset evaluation which proved that our propos ed LSFM generalizes well to unseen data.
****
Self-Supervised Super-Plane for Neural 3D Reconstruction
Botao Ye, Sifei Liu, Xueting Li, Ming-Hsuan Yang; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21415-2142 4
Neural implicit surface representation methods show impressive reconstruction re sults but struggle to handle texture-less planar regions that widely exist in in door scenes. Existing approaches addressing this leverage image prior that requi res assistive networks trained with large-scale annotated datasets. In this work , we introduce a self-supervised super-plane constraint by exploring the free ge ometry cues from the predicted surface, which can further regularize the reconst ruction of plane regions without any other ground truth annotations. Specificall y, we introduce an iterative training scheme, where (i) grouping of pixels to fo

rmulate a super-plane (analogous to super-pixels), and (ii) optimizing of the sc ene reconstruction network via a super-plane constraint, are progressively condu cted. We demonstrate that the model trained with super-planes surprisingly outpe rforms the one using conventional annotated planes, as individual super-plane st atistically occupies a larger area and leads to more stable training. Extensive experiments show that our self-supervised super-plane constraint significantly i mproves 3D reconstruction quality even better than using ground truth plane segm entation. Additionally, the plane reconstruction results from our model can be u sed for auto-labeling for other vision tasks. The code and models are available at https: //github.com/botaoye/S3PRecon.
****
DisCo-CLIP: A Distributed Contrastive Loss for Memory Efficient CLIP Training
Yihao Chen, Xianbiao Qi, Jianan Wang, Lei Zhang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22648-22657
We propose DisCo-CLIP, a distributed memory-efficient CLIP training approach, to reduce the memory consumption of contrastive loss when training contrastive lea rning models. Our approach decomposes the contrastive loss and its gradient comp utation into two parts, one to calculate the intra-GPU gradients and the other t o compute the inter-GPU gradients. According to our decomposition, only the intr a-GPU gradients are computed on the current GPU, while the inter-GPU gradients a re collected via all_reduce from other GPUs instead of being repeatedly computed on every GPU. In this way, we can reduce the GPU memory consumption of contrast ive loss computation from O(B^2) to O(B^2 / N), where B and N are the batch size and the number of GPUs used for training. Such a distributed solution is mathem atically equivalent to the original non-distributed contrastive loss computation , without sacrificing any computation accuracy. It is particularly efficient for large-batch CLIP training. For instance, DisCo-CLIP can enable contrastive trai ning of a ViT-B/32 model with a batch size of 32K or 196K using 8 or 64 A100 40G B GPUs, compared with the original CLIP solution which requires 128 A100 40GB GP Us to train a ViT-B/32 model with a batch size of 32K.
****
GM-NeRF: Learning Generalizable Model-Based Neural Radiance Fields From Multi-Vi ew Images
Jianchuan Chen, Wentao Yi, Liqian Ma, Xu Jia, Huchuan Lu; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20 648-20658
In this work, we focus on synthesizing high-fidelity novel view images for arbit rary human performers, given a set of sparse multi-view images. It is a challeng ing task due to the large variation among articulated body poses and heavy self- occlusions. To alleviate this, we introduce an effective generalizable framework Generalizable Model-based Neural Radiance Fields (GM-NeRF) to synthesize free-v iewpoint images. Specifically, we propose a geometry-guided attention mechanism to register the appearance code from multi-view 2D images to a geometry proxy wh ich can alleviate the misalignment between inaccurate geometry prior and pixel s pace. On top of that, we further conduct neural rendering and partial gradient b ackpropagation for efficient perceptual supervision and improvement of the perce ptual quality of synthesis. To evaluate our method, we conduct experiments on sy nthesized datasets THuman2.0 and Multi-garment, and real-world datasets Genebody and ZJUMocap. The results demonstrate that our approach outperforms state-of-th e-art methods in terms of novel view synthesis and geometric reconstruction.
****
VDN-NeRF: Resolving Shape-Radiance Ambiguity via View-Dependence Normalization
Bingfan Zhu, Yanchao Yang, Xulong Wang, Youyi Zheng, Leonidas Guibas; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 35-45
We propose VDN-NeRF, a method to train neural radiance fields (NeRFs) for better geometry under non-Lambertian surface and dynamic lighting conditions that caus e significant variation in the radiance of a point when viewed from different an gles. Instead of explicitly modeling the underlying factors that result in the v iew-dependent phenomenon, which could be complex yet not inclusive, we develop a

simple and effective technique that normalizes the view-dependence by distilling invariant information already encoded in the learned NeRFs. We then jointly train NeRFs for view synthesis with view-dependence normalization to attain quality geometry. Our experiments show that even though shape-radiance ambiguity is inevitable, the proposed normalization can minimize its effect on geometry, which essentially aligns the optimal capacity needed for explaining view-dependent variations. Our method applies to various baselines and significantly improves geometry without changing the volume rendering pipeline, even if the data is captured under a moving light source. Code is available at: https://github.com/BoifZ/VDN-NeRF.
****
Mobile User Interface Element Detection via Adaptively Prompt Tuning
Zhangxuan Gu, Zhuoer Xu, Haoxing Chen, Jun Lan, Changhua Meng, Weiqiang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11155-11164
Recent object detection approaches rely on pretrained vision-language models for image-text alignment. However, they fail to detect the Mobile User Interface (MUI) element since it contains additional OCR information, which describes its content and function but is often ignored. In this paper, we develop a new MUI element detection dataset named MUI-zh and propose an Adaptively Prompt Tuning (APT) module to take advantage of discriminating OCR information. APT is a lightweight and effective module to jointly optimize category prompts across different modalities. For every element, APT uniformly encodes its visual features and OCR descriptions to dynamically adjust the representation of frozen category prompts. We evaluate the effectiveness of our plug-and-play APT upon several existing CLIP-based detectors for both standard and open-vocabulary MUI element detection. Extensive experiments show that our method achieves considerable improvements on two datasets. The datasets is available at github.com/antmachineintelligence/MUI-zh.
****
Perspective Fields for Single Image Camera Calibration
Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, David F. Fouhey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17307-17316
Geometric camera calibration is often required for applications that understand the perspective of the image. We propose perspective fields as a representation that models the local perspective properties of an image. Perspective Fields contain per-pixel information about the camera view, parameterized as an up vector and a latitude value. This representation has a number of advantages as it makes minimal assumptions about the camera model and is invariant or equivariant to common image editing operations like cropping, warping, and rotation. It is also more interpretable and aligned with human perception. We train a neural network to predict Perspective Fields and the predicted Perspective Fields can be converted to calibration parameters easily. We demonstrate the robustness of our approach under various scenarios compared with camera calibration-based methods and show example applications in image compositing. Project page: https://jinlinyi.github.io/PerspectiveFields/
****
Sparse Multi-Modal Graph Transformer With Shared-Context Processing for Representation Learning of Giga-Pixel Images
Ramin Nakhli, Puria Azadi Moghadam, Haoyang Mi, Hossein Farahani, Alexander Baras, Blake Gilks, Ali Bashashati; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11547-11557
Processing giga-pixel whole slide histopathology images (WSI) is a computationally expensive task. Multiple instance learning (MIL) has become the conventional approach to process WSIs, in which these images are split into smaller patches for further processing. However, MIL-based techniques ignore explicit information about the individual cells within a patch. In this paper, by defining the novel concept of shared-context processing, we designed a multi-modal Graph Transformer that uses the cellular graph within the tissue to provide a single representa

tion for a patient while taking advantage of the hierarchical structure of the tissue, enabling a dynamic focus between cell-level and tissue-level information. We benchmarked the performance of our model against multiple state-of-the-art methods in survival prediction and showed that ours can significantly outperform all of them including hierarchical vision Transformer (ViT). More importantly, we show that our model is strongly robust to missing information to an extent that it can achieve the same performance with as low as 20% of the data. Finally, in two different cancer datasets, we demonstrated that our model was able to stratify the patients into low-risk and high-risk groups while other state-of-the-art methods failed to achieve this goal. We also publish a large dataset of immunohistochemistry (IHC) images containing 1,600 tissue microarray (TMA) cores from 188 patients along with their survival information, making it one of the largest publicly available datasets in this context.
****
Generating Human Motion From Textual Descriptions With Discrete Representations
Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14730-14740
In this work, we investigate a simple and must-known conditional generative framework based on Vector Quantised-Variational AutoEncoder (VQ-VAE) and Generative Pre-trained Transformer (GPT) for human motion generation from textural descriptions. We show that a simple CNN-based VQ-VAE with commonly used training recipes (EMA and Code Reset) allows us to obtain high-quality discrete representations. For GPT, we incorporate a simple corruption strategy during the training to alleviate training-testing discrepancy. Despite its simplicity, our T2M-GPT shows better performance than competitive approaches, including recent diffusion-based approaches. For example, on HumanML3D, which is currently the largest dataset, we achieve comparable performance on the consistency between text and generated motion (R-Precision), but with FID 0.116 largely outperforming MotionDiffuse of 0.630. Additionally, we conduct analyses on HumanML3D and observe that the dataset size is a limitation of our approach. Our work suggests that VQ-VAE still remains a competitive approach for human motion generation. Our implementation is available on the project page: https://mael-zys.github.io/T2M-GPT/
****
Spatial-Temporal Concept Based Explanation of 3D ConvNets
Ying Ji, Yu Wang, Jien Kato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15444-15453
Convolutional neural networks (CNNs) have shown remarkable performance on various tasks. Despite its widespread adoption, the decision procedure of the network still lacks transparency and interpretability, making it difficult to enhance the performance further. Hence, there has been considerable interest in providing explanation and interpretability for CNNs over the last few years. Explainable artificial intelligence (XAI) investigates the relationship between input images or videos and output predictions. Recent studies have achieved outstanding success in explaining 2D image classification ConvNets. On the other hand, due to the high computation cost and complexity of video data, the explanation of 3D video recognition ConvNets is relatively less studied. And none of them are able to produce a high-level explanation. In this paper, we propose a STCE (Spatial-temporal Concept-based Explanation) framework for interpreting 3D ConvNets. In our approach: (1) videos are represented with high-level supervoxels, similar supervoxels are clustered as a concept, which is straightforward for human to understand; and (2) the interpreting framework calculates a score for each concept, which reflects its significance in the ConvNet decision procedure. Experiments on diverse 3D ConvNets demonstrate that our method can identify global concepts with different importance levels, allowing us to investigate the impact of the concepts on a target task, such as action recognition, in-depth. The source codes are publicly available at https://github.com/yingji425/STCE.
****
Robust Test-Time Adaptation in Dynamic Scenarios
Longhui Yuan, Binhui Xie, Shuang Li; Proceedings of the IEEE/CVF Conference on C

omputer Vision and Pattern Recognition (CVPR), 2023, pp. 15922-15932
Test-time adaptation (TTA) intends to adapt the pretrained model to test distributions with only unlabeled test data streams. Most of the previous TTA methods have achieved great success on simple test data streams such as independently sampled data from single or multiple distributions. However, these attempts may fail in dynamic scenarios of real-world applications like autonomous driving, where the environments gradually change and the test data is sampled correlatively over time. In this work, we explore such practical test data streams to deploy the model on the fly, namely practical test-time adaptation (PTTA). To do so, we elaborate a Robust Test-Time Adaptation (RoTTA) method against the complex data stream in PTTA. More specifically, we present a robust batch normalization scheme to estimate the normalization statistics. Meanwhile, a memory bank is utilized to sample category-balanced data with consideration of timeliness and uncertainty. Further, to stabilize the training procedure, we develop a time-aware reweighting strategy with a teacher-student model. Extensive experiments prove that RoTTA enables continual testtime adaptation on the correlatively sampled data streams. Our method is easy to implement, making it a good choice for rapid deployment. The code is publicly available at https://github.com/BIT-DA/RoTTA
****
Global and Local Mixture Consistency Cumulative Learning for Long-Tailed Visual Recognitions
Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, Yun Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15814-15823
In this paper, our goal is to design a simple learning paradigm for long-tail visual recognition, which not only improves the robustness of the feature extractor but also alleviates the bias of the classifier towards head classes while reducing the training skills and overhead. We propose an efficient one-stage training strategy for long-tailed visual recognition called Global and Local Mixture Consistency cumulative learning (GLMC). Our core ideas are twofold: (1) a global and local mixture consistency loss improves the robustness of the feature extractor. Specifically, we generate two augmented batches by the global MixUp and local CutMix from the same batch data, respectively, and then use cosine similarity to minimize the difference. (2) A cumulative head-tail soft label reweighted loss mitigates the head class bias problem. We use empirical class frequencies to reweight the mixed label of the head-tail class for long-tailed data and then balance the conventional loss and the rebalanced loss with a coefficient accumulated by epochs. Our approach achieves state-of-the-art accuracy on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT datasets. Additional experiments on balanced ImageNet and CIFAR demonstrate that GLMC can significantly improve the generalization of backbones. Code is made publicly available at https://github.com/ynu-yangpeng/GLMC
****
NIRVANA: Neural Implicit Representations of Videos With Adaptive Networks and Autoregressive Patch-Wise Modeling
Shishira R. Maiya, Sharath Girish, Max Ehrlich, Hanyu Wang, Kwot Sin Lee, Patrick Poirson, Pengxiang Wu, Chen Wang, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14378-14387
Implicit Neural Representations (INR) have recently shown to be powerful tool for high-quality video compression. However, existing works are limiting as they do not explicitly exploit the temporal redundancy in videos, leading to a long encoding time. Additionally, these methods have fixed architectures which do not scale to longer videos or higher resolutions. To address these issues, we propose NIRVANA, which treats videos as groups of frames and fits separate networks to each group performing patch-wise prediction. %This design shares computation within each group, in the spatial and temporal dimensions, resulting in reduced encoding time of the video. The video representation is modeled autoregressively, with networks fit on a current group initialized using weights from the previous group's model. To further enhance efficiency, we perform quantization of the net

work parameters during training, requiring no post-hoc pruning or quantization. When compared with previous works on the benchmark UVG dataset, NIRVANA improves encoding quality from 37.36 to 37.70 (in terms of PSNR) and the encoding speed by 12x, while maintaining the same compression rate. In contrast to prior video INR works which struggle with larger resolution and longer videos, we show that our algorithm is highly flexible and scales naturally due to its patch-wise and autoregressive designs. Moreover, our method achieves variable bitrate compression by adapting to videos with varying inter-frame motion. NIRVANA achieves 6x decoding speed and scales well with more GPUs, making it practical for various deployment scenarios.
****
Towards Accurate Image Coding: Improved Autoregressive Image Generation With Dynamic Vector Quantization
Mengqi Huang, Zhendong Mao, Zhuowei Chen, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22596-22605
Existing vector quantization (VQ) based autoregressive models follow a two-stage generation paradigm that first learns a codebook to encode images as discrete codes, and then completes generation based on the learned codebook. However, they encode fixed-size image regions into fixed-length codes and ignore their naturally different information densities, which results in insufficiency in important regions and redundancy in unimportant ones, and finally degrades the generation quality and speed. Moreover, the fixed-length coding leads to an unnatural raster-scan autoregressive generation. To address the problem, we propose a novel two-stage framework: (1) Dynamic-Quantization VAE (DQ-VAE) which encodes image regions into variable-length codes based on their information densities for an accurate & compact code representation. (2) DQ-Transformer which thereby generates images autoregressively from coarse-grained (smooth regions with fewer codes) to fine-grained (details regions with more codes) by modeling the position and content of codes in each granularity alternately, through a novel stacked-transformer architecture and shared-content, non-shared position input layers designs. Comprehensive experiments on various generation tasks validate our superiorities in both effectiveness and efficiency.
****
Coaching a Teachable Student
Jimuyang Zhang, Zanming Huang, Eshed Ohn-Bar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7805-7815
We propose a novel knowledge distillation framework for effectively teaching a sensorimotor student agent to drive from the supervision of a privileged teacher agent. Current distillation for sensorimotor agents methods tend to result in suboptimal learned driving behavior by the student, which we hypothesize is due to inherent differences between the input, modeling capacity, and optimization processes of the two agents. We develop a novel distillation scheme that can address these limitations and close the gap between the sensorimotor agent and its privileged teacher. Our key insight is to design a student which learns to align their input features with the teacher's privileged Bird's Eye View (BEV) space. The student then can benefit from direct supervision by the teacher over the internal representation learning. To scaffold the difficult sensorimotor learning task, the student model is optimized via a student-paced coaching mechanism with various auxiliary supervision. We further propose a high-capacity imitation learned privileged agent that surpasses prior privileged agents in CARLA and ensures the student learns safe driving behavior. Our proposed sensorimotor agent results in a robust image-based behavior cloning agent in CARLA, improving over current models by over 20.6% in driving score without requiring LiDAR, historical observations, ensemble of models, on-policy data aggregation or reinforcement learning.
****
Collaboration Helps Camera Overtake LiDAR in 3D Detection
Yue Hu, Yifan Lu, Runsheng Xu, Weidi Xie, Siheng Chen, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2

023, pp. 9243-9252
Camera-only 3D detection provides an economical solution with a simple configura tion for localizing objects in 3D space compared to LiDAR-based detection system s. However, a major challenge lies in precise depth estimation due to the lack o f direct 3D measurements in the input. Many previous methods attempt to improve depth estimation through network designs, e.g., deformable layers and larger rec eptive fields. This work proposes an orthogonal direction, improving the camera-only 3D detection by introducing multi-agent collaborations. Our proposed collab orative camera-only 3D detection (CoCa3D) enables agents to share complementary information with each other through communication. Meanwhile, we optimize commun ication efficiency by selecting the most informative cues. The shared messages f rom multiple viewpoints disambiguate the single-agent estimated depth and comple ment the occluded and long-range regions in the single-agent view. We evaluate C oCa3D in one real-world dataset and two new simulation datasets. Results show th at CoCa3D improves previous SOTA performances by 44.21% on DAIR-V2X, 30.60% on O PV2V+, 12.59% on CoPerception-UAVs+ for AP@70. Our preliminary results show a po tential that with sufficient collaboration, the camera might overtake LiDAR in s ome practical scenarios. We released the dataset and code at https://siheng-chen .github.io/dataset/CoPerception+ and https://github.com/MediaBrain-SJTU/CoCa3D.
****
RealImpact: A Dataset of Impact Sound Fields for Real Objects
Samuel Clarke, Ruohan Gao, Mason Wang, Mark Rau, Julia Xu, Jui-Hsien Wang, Doug L. James, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2023, pp. 1516-1525
Objects make unique sounds under different perturbations, environment conditions , and poses relative to the listener. While prior works have modeled impact soun ds and sound propagation in simulation, we lack a standard dataset of impact sou nd fields of real objects for audio-visual learning and calibration of the sim-t o-real gap. We present RealImpact, a large-scale dataset of real object impact s ounds recorded under controlled conditions. RealImpact contains 150,000 recordin gs of impact sounds of 50 everyday objects with detailed annotations, including their impact locations, microphone locations, contact force profiles, material l abels, and RGBD images. We make preliminary attempts to use our dataset as a ref erence to current simulation methods for estimating object impact sounds that ma tch the real world. Moreover, we demonstrate the usefulness of our dataset as a testbed for acoustic and audio-visual learning via the evaluation of two benchma rk tasks, including listener location classification and visual acoustic matchin g.
****
ReCo: Region-Controlled Text-to-Image Generation
Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Du an, Zicheng Liu, Ce Liu, Michael Zeng, Lijuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14246-14 255
Recently, large-scale text-to-image (T2I) models have shown impressive performan ce in generating high-fidelity images, but with limited controllability, e.g., p recisely specifying the content in a specific region with a free-form text descr iption. In this paper, we propose an effective technique for such regional contr ol in T2I generation. We augment T2I models' inputs with an extra set of positio n tokens, which represent the quantized spatial coordinates. Each region is spec ified by four position tokens to represent the top-left and bottom-right corners , followed by an open-ended natural language regional description. Then, we fine -tune a pre-trained T2I model with such new input interface. Our model, dubbed a s ReCo (Region-Controlled T2I), enables the region control for arbitrary objects described by open-ended regional texts rather than by object labels from a cons trained category set. Empirically, ReCo achieves better image quality than the T 2I model strengthened by positional words (FID: 8.82 -> 7.36, SceneFID: 15.54 -> 6.51 on COCO), together with objects being more accurately placed, amounting to a 20.40% region classification accuracy improvement on COCO. Furthermore, we de monstrate that ReCo can better control the object count, spatial relationship, a

nd region attributes such as color/size, with the free-form regional description. Human evaluation on PaintSkill shows that ReCo is +19.28% and +17.21% more accurate in generating images with correct object count and spatial relationship than the T2I model.
****
WINNER: Weakly-Supervised hIerarchical decompositioN and aligNment for Spatio-tEmporal Video gRounding
Mengze Li, Han Wang, Wenqiao Zhang, Jiaxu Miao, Zhou Zhao, Shengyu Zhang, Wei Ji, Fei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23090-23099
Spatio-temporal video grounding aims to localize the aligned visual tube corresponding to a language query. Existing techniques achieve such alignment by exploiting dense boundary and bounding box annotations, which can be prohibitively expensive. To bridge the gap, we investigate the weakly-supervised setting, where models learn from easily accessible video-language data without annotations. We identify that intra-sample spurious correlations among video-language components can be alleviated if the model captures the decomposed structures of video and language data. In this light, we propose a novel framework, namely WINNER, for hierarchical video-text understanding. WINNER first builds the language decomposition tree in a bottom-up manner, upon which the structural attention mechanism and top-down feature backtracking jointly build a multi-modal decomposition tree, permitting a hierarchical understanding of unstructured videos. The multi-modal decomposition tree serves as the basis for multi-hierarchy language-tube matching. A hierarchical contrastive learning objective is proposed to learn the multi-hierarchy correspondence and distinguishment with intra-sample and inter-sample video-text decomposition structures, achieving video-language decomposition structure alignment. Extensive experiments demonstrate the rationality of our design and its effectiveness beyond state-of-the-art weakly supervised methods, even some supervised methods.
****
Preserving Linear Separability in Continual Learning by Backward Feature Projection
Qiao Gu, Dongsub Shim, Florian Shkurti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24286-24295
Catastrophic forgetting has been a major challenge in continual learning, where the model needs to learn new tasks with limited or no access to data from previously seen tasks. To tackle this challenge, methods based on knowledge distillation in feature space have been proposed and shown to reduce forgetting. However, most feature distillation methods directly constrain the new features to match the old ones, overlooking the need for plasticity. To achieve a better stability-plasticity trade-off, we propose Backward Feature Projection (BFP), a method for continual learning that allows the new features to change up to a learnable linear transformation of the old features. BFP preserves the linear separability of the old classes while allowing the emergence of new feature directions to accommodate new classes. BFP can be integrated with existing experience replay methods and boost performance by a significant margin. We also demonstrate that BFP helps learn a better representation space, in which linear separability is well preserved during continual learning and linear probing achieves high classification accuracy.
****
MHPL: Minimum Happy Points Learning for Active Source Free Domain Adaptation
Fan Wang, Zhongyi Han, Zhiyan Zhang, Rundong He, Yilong Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20008-20018
Source free domain adaptation (SFDA) aims to transfer a trained source model to the unlabeled target domain without accessing the source data. However, the SFDA setting faces a performance bottleneck due to the absence of source data and target supervised information, as evidenced by the limited performance gains of the newest SFDA methods. Active source free domain adaptation (ASFDA) can break through the problem by exploring and exploiting a small set of informative samples

via active learning. In this paper, we first find that those satisfying the properties of neighbor-chaotic, individual-different, and source-dissimilar are the best points to select. We define them as the minimum happy (MH) points challenging to explore with existing methods. We propose minimum happy points learning (MHPL) to explore and exploit MH points actively. We design three unique strategies: neighbor environment uncertainty, neighbor diversity relaxation, and one-shot querying, to explore the MH points. Further, to fully exploit MH points in the learning process, we design a neighbor focal loss that assigns the weighted neighbor purity to the cross entropy loss of MH points to make the model focus more on them. Extensive experiments verify that MHPL remarkably exceeds the various types of baselines and achieves significant performance gains at a small cost of labeling.
****
Fix the Noise: Disentangling Source Feature for Controllable Domain Translation
Dongyeun Lee, Jae Young Lee, Doyeon Kim, Jaehyun Choi, Jaejun Yoo, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14224-14234
Recent studies show strong generative performance in domain translation especially by using transfer learning techniques on the unconditional generator. However, the control between different domain features using a single model is still challenging. Existing methods often require additional models, which is computationally demanding and leads to unsatisfactory visual quality. In addition, they have restricted control steps, which prevents a smooth transition. In this paper, we propose a new approach for high-quality domain translation with better controllability. The key idea is to preserve source features within a disentangled subspace of a target feature space. This allows our method to smoothly control the degree to which it preserves source features while generating images from an entirely new domain using only a single model. Our extensive experiments show that the proposed method can produce more consistent and realistic images than previous works and maintain precise controllability over different levels of transformation. The code is available at LeeDongYeun/FixNoise.
****
Metadata-Based RAW Reconstruction via Implicit Neural Functions
Leyi Li, Huijie Qiao, Qi Ye, Qinmin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18196-18205
Many low-level computer vision tasks are desirable to utilize the unprocessed RAW image as input, which remains the linear relationship between pixel values and scene radiance. Recent works advocate to embed the RAW image samples into sRGB images at capture time, and reconstruct the RAW from sRGB by these metadata when needed. However, there still exist some limitations on taking full use of the metadata. In this paper, instead of following the perspective of sRGB-to-RAW mapping, we reformulate the problem as mapping the 2D coordinates of the metadata to its RAW values conditioned on the corresponding sRGB values. With this novel formulation, we propose to reconstruct the RAW image with an implicit neural function, which achieves significant performance improvement (more than 10dB average PSNR) only with the uniform sampling. Compared with most deep learning-based approaches, our method is trained in a self-supervised way that requiring no pre-training on different camera ISPs. We perform further experiments to demonstrate the effectiveness of our method, and show that our framework is also suitable for the task of guided super-resolution.
****
Uni-Perceiver v2: A Generalist Model for Large-Scale Vision and Vision-Language Tasks
Hao Li, Jinguo Zhu, Xiaohu Jiang, Xizhou Zhu, Hongsheng Li, Chun Yuan, Xiaohua Wang, Yu Qiao, Xiaogang Wang, Wenhai Wang, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2691-2700
Despite the remarkable success of foundation models, their task-specific fine-tuning paradigm makes them inconsistent with the goal of general perception modeling. The key to eliminating this inconsistency is to use generalist models for ge

neral task modeling. However, existing attempts at generalist models are inadequate in both versatility and performance. In this paper, we propose Uni-Perceiver v2, which is the first generalist model capable of handling major large-scale vision and vision-language tasks with competitive performance. Specifically, images are encoded as general region proposals, while texts are encoded via a Transformer-based language model. The encoded representations are transformed by a task-agnostic decoder. Different tasks are formulated as a unified maximum likelihood estimation problem. We further propose an effective optimization technique named Task-Balanced Gradient Normalization to ensure stable multi-task learning with an unmixed sampling strategy, which is helpful for tasks requiring large batch-size training. After being jointly trained on various tasks, Uni-Perceiver v2 is capable of directly handling downstream tasks without any task-specific adaptation. Results show that Uni-Perceiver v2 outperforms all existing generalist models in both versatility and performance. Meanwhile, compared with the commonly-recognized strong baselines that require tasks-specific fine-tuning, Uni-Perceiver v2 achieves competitive performance on a broad range of vision and vision-language tasks.
****
Sparsely Annotated Semantic Segmentation With Adaptive Gaussian Mixtures
Linshan Wu, Zhun Zhong, Leyuan Fang, Xingxin He, Qiang Liu, Jiayi Ma, Hao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15454-15464
Sparsely annotated semantic segmentation (SASS) aims to learn a segmentation model by images with sparse labels (i.e., points or scribbles). Existing methods mainly focus on introducing low-level affinity or generating pseudo labels to strengthen supervision, while largely ignoring the inherent relation between labeled and unlabeled pixels. In this paper, we observe that pixels that are close to each other in the feature space are more likely to share the same class. Inspired by this, we propose a novel SASS framework, which is equipped with an Adaptive Gaussian Mixture Model (AGMM). Our AGMM can effectively endow reliable supervision for unlabeled pixels based on the distributions of labeled and unlabeled pixels. Specifically, we first build Gaussian mixtures using labeled pixels and their relatively similar unlabeled pixels, where the labeled pixels act as centroids, for modeling the feature distribution of each class. Then, we leverage the reliable information from labeled pixels and adaptively generated GMM predictions to supervise the training of unlabeled pixels, achieving online, dynamic, and robust self-supervision. In addition, by capturing category-wise Gaussian mixtures, AGMM encourages the model to learn discriminative class decision boundaries in an end-to-end contrastive learning manner. Experimental results conducted on the PASCAL VOC 2012 and Cityscapes datasets demonstrate that our AGMM can establish new state-of-the-art SASS performance. Code is available at https://github.com/Luffy03/AGMM-SASS.
****
Multimodality Helps Unimodality: Cross-Modal Few-Shot Learning With Multimodal Models
Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19325-19337
The ability to quickly learn a new task with minimal instruction - known as few-shot learning - is a central aspect of intelligent agents. Classical few-shot benchmarks make use of few-shot samples from a single modality, but such samples may not be sufficient to characterize an entire concept class. In contrast, humans use cross-modal information to learn new concepts efficiently. In this work, we demonstrate that one can indeed build a better visual dog classifier by reading about dogs and listening to them bark. To do so, we exploit the fact that recent multimodal foundation models such as CLIP are inherently cross-modal, mapping different modalities to the same representation space. Specifically, we propose a simple cross-modal adaptation approach that learns from few-shot examples spanning different modalities. By repurposing class names as additional one-shot training samples, we achieve SOTA results with an embarrassingly simple linear cla

ssifier for vision-language adaptation. Furthermore, we show that our approach can benefit existing methods such as prefix tuning and classifier ensembling. Finally, to explore other modalities beyond vision and language, we construct the first (to our knowledge) audiovisual few-shot benchmark and use cross-modal training to improve the performance of both image and audio classification. We hope our success can inspire future works to embrace cross-modality for even broader domains and tasks.
****
Decompose More and Aggregate Better: Two Closer Looks at Frequency Representation Learning for Human Motion Prediction
Xuehao Gao, Shaoyi Du, Yang Wu, Yang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6451-6460
Encouraged by the effectiveness of encoding temporal dynamics within the frequency domain, recent human motion prediction systems prefer to first convert the motion representation from the original pose space into the frequency space. In this paper, we introduce two closer looks at effective frequency representation learning for robust motion prediction and summarize them as: decompose more and aggregate better. Motivated by these two insights, we develop two powerful units that factorize the frequency representation learning task with a novel decomposition-aggregation two-stage strategy: (1) frequency decomposition unit unweaves multi-view frequency representations from an input body motion by embedding its frequency features into multiple spaces; (2) feature aggregation unit deploys a series of intra-space and inter-space feature aggregation layers to collect comprehensive frequency representations from these spaces for robust human motion prediction. As evaluated on large-scale datasets, we develop a strong baseline model for the human motion prediction task that outperforms state-of-the-art methods by large margins: 8% 12% on Human3.6M, 3% 7% on CMU MoCap, and 7% 10% on 3DPW.
****
Diversity-Aware Meta Visual Prompting
Qidong Huang, Xiaoyi Dong, Dongdong Chen, Weiming Zhang, Feifei Wang, Gang Hua, Nenghai Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10878-10887
We present Diversity-Aware Meta Visual Prompting (DAM-VP), an efficient and effective prompting method for transferring pre-trained models to downstream tasks with frozen backbone. A challenging issue in visual prompting is that image datasets sometimes have a large data diversity whereas a per-dataset generic prompt can hardly handle the complex distribution shift toward the original pretraining data distribution properly. To address this issue, we propose a dataset Diversity-Aware prompting strategy whose initialization is realized by a Meta-prompt. Specifically, we cluster the downstream dataset into small homogeneity subsets in a diversity-adaptive way, with each subset has its own prompt optimized separately. Such a divide-and-conquer design reduces the optimization difficulty greatly and significantly boosts the prompting performance. Furthermore, all the prompts are initialized with a meta-prompt, which is learned across several datasets. It is a bootstrapped paradigm, with the key observation that the prompting knowledge learned from previous datasets could help the prompt to converge faster and perform better on a new dataset. During inference, we dynamically select a proper prompt for each input, based on the feature distance between the input and each subset. Through extensive experiments, our DAM-VP demonstrates superior efficiency and effectiveness, clearly surpassing previous prompting methods in a series of downstream datasets for different pretraining models. Our code is available at: https://github.com/shikiw/DAM-VP.
****
Affection: Learning Affective Explanations for Real-World Visual Data
Panos Achlioptas, Maks Ovsjanikov, Leonidas Guibas, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6641-6651
In this work, we explore the space of emotional reactions induced by real-world images. For this, we first introduce a large-scale dataset that contains both categorical emotional reactions and free-form textual explanations for 85,007 publ

icly available images, analyzed by 6,283 annotators who were asked to indicate and explain how and why they felt when observing a particular image, with a total of 526,749 responses. Although emotional reactions are subjective and sensitive to context (personal mood, social status, past experiences) -- we show that there is significant common ground to capture emotional responses with a large support in the subject population. In light of this observation, we ask the following questions: i) Can we develop neural networks that provide plausible affective responses to real-world visual data explained with language? ii) Can we steer such methods towards producing explanations with varying degrees of pragmatic language, justifying different emotional reactions by grounding them in the visual stimulus? Finally, iii) How to evaluate the performance of such methods for this novel task? In this work, we take the first steps in addressing all of these questions, paving the way for more human-centric and emotionally-aware image analysis systems. Our code and data are publicly available at https://affective-explanations.org.
****
3D Highlighter: Localizing Regions on 3D Shapes via Text Descriptions
Dale Decatur, Itai Lang, Rana Hanocka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20930-20939
We present 3D Highlighter, a technique for localizing semantic regions on a mesh using text as input. A key feature of our system is the ability to interpret "out-of-domain" localizations. Our system demonstrates the ability to reason about where to place non-obviously related concepts on an input 3D shape, such as adding clothing to a bare 3D animal model. Our method contextualizes the text description using a neural field and colors the corresponding region of the shape using a probability-weighted blend. Our neural optimization is guided by a pre-trained CLIP encoder, which bypasses the need for any 3D datasets or 3D annotations. Thus, 3D Highlighter is highly flexible, general, and capable of producing localizations on a myriad of input shapes.
****
Iterative Geometry Encoding Volume for Stereo Matching
Gangwei Xu, Xianqi Wang, Xiaohuan Ding, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21919-21928
Recurrent All-Pairs Field Transforms (RAFT) has shown great potentials in matching tasks. However, all-pairs correlations lack non-local geometry knowledge and have difficulties tackling local ambiguities in ill-posed regions. In this paper, we propose Iterative Geometry Encoding Volume (IGEV-Stereo), a new deep network architecture for stereo matching. The proposed IGEV-Stereo builds a combined geometry encoding volume that encodes geometry and context information as well as local matching details, and iteratively indexes it to update the disparity map. To speed up the convergence, we exploit GEV to regress an accurate starting point for ConvGRUs iterations. Our IGEV-Stereo ranks first on KITTI 2015 and 2012 (Reflective) among all published methods and is the fastest among the top 10 methods. In addition, IGEV-Stereo has strong cross-dataset generalization as well as high inference efficiency. We also extend our IGEV to multi-view stereo (MVS), i.e. IGEV-MVS, which achieves competitive accuracy on DTU benchmark. Code is available at https://github.com/gangweiX/IGEV.
****
PLA: Language-Driven Open-Vocabulary 3D Scene Understanding
Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7010-7019
Open-vocabulary scene understanding aims to localize and recognize unseen categories beyond the annotated label space. The recent breakthrough of 2D open-vocabulary perception is largely driven by Internet-scale paired image-text data with rich vocabulary concepts. However, this success cannot be directly transferred to 3D scenarios due to the inaccessibility of large-scale 3D-text pairs. To this end, we propose to distill knowledge encoded in pre-trained vision-language (VL) foundation models through captioning multi-view images from 3D, which allows ex

plicitly associating 3D and semantic-rich captions. Further, to foster coarse-to-fine visual-semantic representation learning from captions, we design hierarchical 3D-caption pairs, leveraging geometric constraints between 3D scenes and multi-view images. Finally, by employing contrastive learning, the model learns language-aware embeddings that connect 3D and text for open-vocabulary tasks. Our method not only remarkably outperforms baseline methods by 25.8%   44.7% hIoU and 14.5%   50.4% hAP_ 50 in open-vocabulary semantic and instance segmentation, but also shows robust transferability on challenging zero-shot domain transfer tasks. See the project website at https://dingry.github.io/projects/PLA.
****
FaceLit: Neural 3D Relightable Faces
Anurag Ranjan, Kwang Moo Yi, Jen-Hao Rick Chang, Oncel Tuzel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8619-8628
We propose a generative framework, FaceLit, capable of generating a 3D face that can be rendered at various user-defined lighting conditions and views, learned purely from 2D images in-the-wild without any manual annotation. Unlike existing works that require careful capture setup or human labor, we rely on off-the-shelf pose and illumination estimators. With these estimates, we incorporate the Phong reflectance model in the neural volume rendering framework. Our model learns to generate shape and material properties of a face such that, when rendered according to the natural statistics of pose and illumination, produces photorealistic face images with multiview 3D and illumination consistency. Our method enables photorealistic generation of faces with explicit illumination and view controls on multiple datasets -- FFHQ, MetFaces and CelebA-HQ. We show state-of-the-art photorealism among 3D aware GANs on FFHQ dataset achieving an FID score of 3.5.
****
Visual Programming: Compositional Visual Reasoning Without Training
Tanmay Gupta, Aniruddha Kembhavi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14953-14962
We present VISPROG, a neuro-symbolic approach to solving complex and compositional visual tasks given natural language instructions. VISPROG avoids the need for any task-specific training. Instead, it uses the in-context learning ability of large language models to generate python-like modular programs, which are then executed to get both the solution and a comprehensive and interpretable rationale. Each line of the generated program may invoke one of several off-the-shelf computer vision models, image processing routines, or python functions to produce intermediate outputs that may be consumed by subsequent parts of the program. We demonstrate the flexibility of VISPROG on 4 diverse tasks - compositional visual question answering, zero-shot reasoning on image pairs, factual knowledge object tagging, and language-guided image editing. We believe neuro-symbolic approaches like VISPROG are an exciting avenue to easily and effectively expand the scope of AI systems to serve the long tail of complex tasks that people may wish to perform.
****
InstMove: Instance Motion for Object-Centric Video Segmentation
Qihao Liu, Junfeng Wu, Yi Jiang, Xiang Bai, Alan L. Yuille, Song Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6344-6354
Despite significant efforts, cutting-edge video segmentation methods still remain sensitive to occlusion and rapid movement, due to their reliance on the appearance of objects in the form of object embeddings, which are vulnerable to these disturbances. A common solution is to use optical flow to provide motion information, but essentially it only considers pixel-level motion, which still relies on appearance similarity and hence is often inaccurate under occlusion and fast movement. In this work, we study the instance-level motion and present InstMove, which stands for Instance Motion for Object-centric Video Segmentation. In comparison to pixel-wise motion, InstMove mainly relies on instance-level motion information that is free from image feature embeddings, and features physical interp

retations, making it more accurate and robust toward occlusion and fast-moving o
bjects. To better fit in with the video segmentation tasks, InstMove uses instan
ce masks to model the physical presence of an object and learns the dynamic mode
l through a memory network to predict its position and shape in the next frame.
With only a few lines of code, InstMove can be integrated into current SOTA meth
ods for three different video segmentation tasks and boost their performance. Sp
ecifically, we improve the previous arts by 1.5 AP on OVIS dataset, which featur
es heavy occlusions, and 4.9 AP on YouTubeVIS-Long dataset, which mainly contain
s fast-moving objects. These results suggest that instance-level motion is robus
t and accurate, and hence serving as a powerful solution in complex scenarios fo
r object-centric video segmentation.
****
Real-Time Evaluation in Online Continual Learning: A New Hope
Yasir Ghunaim, Adel Bibi, Kumail Alhamoud, Motasem Alfarra, Hasan Abed Al Kader
Hammoud, Ameya Prabhu, Philip H.S. Torr, Bernard Ghanem; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 118
88-11897
Current evaluations of Continual Learning (CL) methods typically assume that the
re is no constraint on training time and computation. This is an unrealistic ass
umption for any real-world setting, which motivates us to propose: a practical r
eal-time evaluation of continual learning, in which the stream does not wait for
the model to complete training before revealing the next data for predictions.
To do this, we evaluate current CL methods with respect to their computational c
osts. We conduct extensive experiments on CLOC, a large-scale dataset containing
39 million time-stamped images with geolocation labels. We show that a simple b
aseline outperforms state-of-the-art CL methods under this evaluation, questioni
ng the applicability of existing methods in realistic settings. In addition, we
explore various CL components commonly used in the literature, including memory
sampling strategies and regularization approaches. We find that all considered m
ethods fail to be competitive against our simple baseline. This surprisingly sug
gests that the majority of existing CL literature is tailored to a specific clas
s of streams that is not practical. We hope that the evaluation we provide will
be the first step towards a paradigm shift to consider the computational cost in
the development of online continual learning methods.
****
GRES: Generalized Referring Expression Segmentation
Chang Liu, Henghui Ding, Xudong Jiang; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23592-23601
Referring Expression Segmentation (RES) aims to generate a segmentation mask for
the object described by a given language expression. Existing classic RES datas
ets and methods commonly support single-target expressions only, i.e., one expre
ssion refers to one target object. Multi-target and no-target expressions are no
t considered. This limits the usage of RES in practice. In this paper, we introd
uce a new benchmark called Generalized Referring Expression Segmentation (GRES),
which extends the classic RES to allow expressions to refer to an arbitrary num
ber of target objects. Towards this, we construct the first large-scale GRES dat
aset called gRefCOCO that contains multi-target, no-target, and single-target ex
pressions. GRES and gRefCOCO are designed to be well-compatible with RES, facili
tating extensive experiments to study the performance gap of the existing RES me
thods on the GRES task. In the experimental study, we find that one of the big c
hallenges of GRES is complex relationship modeling. Based on this, we propose a
region-based GRES baseline ReLA that adaptively divides the image into regions w
ith sub-instance clues, and explicitly models the region-region and region-langu
age dependencies. The proposed approach ReLA achieves new state-of-the-art perfo
rmance on the both newly proposed GRES and classic RES tasks. The proposed gRefC
OCO dataset and method are available at https://henghuiding.github.io/GRES.
****
Towards Effective Adversarial Textured 3D Meshes on Physical Face Recognition
Xiao Yang, Chang Liu, Longlong Xu, Yikai Wang, Yinpeng Dong, Ning Chen, Hang Su,
Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2023, pp. 4119-4128
Face recognition is a prevailing authentication solution in numerous biometric applications. Physical adversarial attacks, as an important surrogate, can identify the weaknesses of face recognition systems and evaluate their robustness before deployed. However, most existing physical attacks are either detectable readily or ineffective against commercial recognition systems. The goal of this work is to develop a more reliable technique that can carry out an end-to-end evaluation of adversarial robustness for commercial systems. It requires that this technique can simultaneously deceive black-box recognition models and evade defensive mechanisms. To fulfill this, we design adversarial textured 3D meshes (AT3D) with an elaborate topology on a human face, which can be 3D-printed and pasted on the attacker's face to evade the defenses. However, the mesh-based optimization regime calculates gradients in high-dimensional mesh space, and can be trapped into local optima with unsatisfactory transferability. To deviate from the mesh-based space, we propose to perturb the low-dimensional coefficient space based on 3D Morphable Model, which significantly improves black-box transferability meanwhile enjoying faster search efficiency and better visual quality. Extensive experiments in digital and physical scenarios show that our method effectively explores the security vulnerabilities of multiple popular commercial services, including three recognition APIs, four anti-spoofing APIs, two prevailing mobile phones and two automated access control systems.
****
BAAM: Monocular 3D Pose and Shape Reconstruction With Bi-Contextual Attention Module and Attention-Guided Modeling
Hyo-Jun Lee, Hanul Kim, Su-Min Choi, Seong-Gyun Jeong, Yeong Jun Koh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9011-9020
3D traffic scene comprises various 3D information about car objects, including their pose and shape. However, most recent studies pay relatively less attention to reconstructing detailed shapes. Furthermore, most of them treat each 3D object as an independent one, resulting in losses of relative context inter-objects and scene context reflecting road circumstances. A novel monocular 3D pose and shape reconstruction algorithm, based on bi-contextual attention and attention-guided modeling (BAAM), is proposed in this work. First, given 2D primitives, we reconstruct 3D object shape based on attention-guided modeling that considers the relevance between detected objects and vehicle shape priors. Next, we estimate 3D object pose through bi-contextual attention, which leverages relation-context inter objects and scene-context between an object and road environment. Finally, we propose a 3D non maximum suppression algorithm to eliminate spurious objects based on their Bird-Eye-View distance. Extensive experiments demonstrate that the proposed BAAM yields state-of-the-art performance on ApolloCar3D. Also, they show that the proposed BAAM can be plugged into any mature monocular 3D object detector on KITTI and significantly boost their performance.
****
Freestyle Layout-to-Image Synthesis
Han Xue, Zhiwu Huang, Qianru Sun, Li Song, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14256-14266
Typical layout-to-image synthesis (LIS) models generate images for a closed set of semantic classes, e.g., 182 common objects in COCO-Stuff. In this work, we explore the freestyle capability of the model, i.e., how far can it generate unseen semantics (e.g., classes, attributes, and styles) onto a given layout, and call the task Freestyle LIS (FLIS). Thanks to the development of large-scale pre-trained language-image models, a number of discriminative models (e.g., image classification and object detection) trained on limited base classes are empowered with the ability of unseen class prediction. Inspired by this, we opt to leverage large-scale pre-trained text-to-image diffusion models to achieve the generation of unseen semantics. The key challenge of FLIS is how to enable the diffusion model to synthesize images from a specific layout which very likely violates its pre-learned knowledge, e.g., the model never sees "a unicorn sitting on a bench

" during its pre-training. To this end, we introduce a new module called Rectified Cross-Attention (RCA) that can be conveniently plugged in the diffusion model to integrate semantic masks. This "plug-in" is applied in each cross-attention layer of the model to rectify the attention maps between image and text tokens. The key idea of RCA is to enforce each text token to act on the pixels in a specified region, allowing us to freely put a wide variety of semantics from pre-trained knowledge (which is general) onto the given layout (which is specific). Extensive experiments show that the proposed diffusion network produces realistic and freestyle layout-to-image generation results with diverse text inputs, which has a high potential to spawn a bunch of interesting applications. Code is available at https://github.com/essunny310/FreestyleNet.
****
Effective Ambiguity Attack Against Passport-Based DNN Intellectual Property Protection Schemes Through Fully Connected Layer Substitution
Yiming Chen, Jinyu Tian, Xiangyu Chen, Jiantao Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8123-8132
Since training a deep neural network (DNN) is costly, the well-trained deep models can be regarded as valuable intellectual property (IP) assets. The IP protection associated with deep models has been receiving increasing attentions in recent years. Passport-based method, which replaces normalization layers with passport layers, has been one of the few protection solutions that are claimed to be secure against advanced attacks. In this work, we tackle the issue of evaluating the security of passport-based IP protection methods. We propose a novel and effective ambiguity attack against passport-based method, capable of successfully forging multiple valid passports with a small training dataset. This is accomplished by inserting a specially designed accessory block ahead of the passport parameters. Using less than 10% of training data, with the forged passport, the model exhibits almost indistinguishable performance difference (less than 2%) compared with that of the authorized passport. In addition, it is shown that our attack strategy can be readily generalized to attack other IP protection methods based on watermark embedding. Directions for potential remedy solutions are also given.
****
Visual Dependency Transformers: Dependency Tree Emerges From Reversed Attention
Mingyu Ding, Yikang Shen, Lijie Fan, Zhenfang Chen, Zitian Chen, Ping Luo, Joshua B. Tenenbaum, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14528-14539
Humans possess a versatile mechanism for extracting structured representations of our visual world. When looking at an image, we can decompose the scene into entities and their parts as well as obtain the dependencies between them. To mimic such capability, we propose Visual Dependency Transformers (DependencyViT) that can induce visual dependencies without any labels. We achieve that with a novel neural operator called reversed attention that can naturally capture long-range visual dependencies between image patches. Specifically, we formulate it as a dependency graph where a child token in reversed attention is trained to attend to its parent tokens and send information following a normalized probability distribution rather than gathering information in conventional self-attention. With such a design, hierarchies naturally emerge from reversed attention layers, and a dependency tree is progressively induced from leaf nodes to the root node unsupervisedly. DependencyViT offers several appealing benefits. (i) Entities and their parts in an image are represented by different subtrees, enabling part partitioning from dependencies; (ii) Dynamic visual pooling is made possible. The leaf nodes which rarely send messages can be pruned without hindering the model performance, based on which we propose the lightweight DependencyViT-Lite to reduce the computational and memory footprints; (iii) DependencyViT works well on both self- and weakly-supervised pretraining paradigms on ImageNet, and demonstrates its effectiveness on 8 datasets and 5 tasks, such as unsupervised part and saliency segmentation, recognition, and detection.
****

Differentiable Architecture Search With Random Features
Xuanyang Zhang, Yonggang Li, Xiangyu Zhang, Yongtao Wang, Jian Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16060-16069
Differentiable architecture search (DARTS) has significantly promoted the development of NAS techniques because of its high search efficiency and effectiveness but suffers from performance collapse. In this paper, we make efforts to alleviate the performance collapse problem for DARTS from two aspects. First, we investigate the expressive power of the supernet in DARTS and then derive a new setup of DARTS paradigm with only training BatchNorm. Second, we theoretically find that random features dilute the auxiliary connection role of skip-connection in supernet optimization and enable search algorithm focus on fairer operation selection, thereby solving the performance collapse problem. We instantiate DARTS and PC-DARTS with random features to build an improved version for each named RF-DARTS and RF-PCDARTS respectively. Experimental results show that RF-DARTS obtains 94.36% test accuracy on CIFAR-10 (which is the nearest optimal result in NAS-Bench-201), and achieves the newest state-of-the-art top-1 test error of 24.0% on ImageNet when transferring from CIFAR-10. Moreover, RF-DARTS performs robustly across three datasets (CIFAR-10, CIFAR-100, and SVHN) and four search spaces (S1-S4). Besides, RF-PCDARTS achieves even better results on ImageNet, that is, 23.9% top-1 and 7.1% top-5 test error, surpassing representative methods like single-path, training-free, and partial-channel paradigms directly searched on ImageNet.
****
Open-Set Fine-Grained Retrieval via Prompting Vision-Language Evaluator
Shijie Wang, Jianlong Chang, Haojie Li, Zhihui Wang, Wanli Ouyang, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19381-19391
Open-set fine-grained retrieval is an emerging challenge that requires an extra capability to retrieve unknown subcategories during evaluation. However, current works are rooted in the close-set scenarios, where all the subcategories are pre-defined, and make it hard to capture discriminative knowledge from unknown subcategories, consequently failing to handle the inevitable unknown subcategories in open-world scenarios. In this work, we propose a novel Prompting vision-Language Evaluator (PLEor) framework based on the recently introduced contrastive language-image pretraining (CLIP) model, for open-set fine-grained retrieval. PLEor could leverage pre-trained CLIP model to infer the discrepancies encompassing both pre-defined and unknown subcategories, called category-specific discrepancies, and transfer them to the backbone network trained in the close-set scenarios. To make pre-trained CLIP model sensitive to category-specific discrepancies, we design a dual prompt scheme to learn a vision prompt specifying the category-specific discrepancies, and turn random vectors with category names in a text prompt into category-specific discrepancy descriptions. Moreover, a vision-language evaluator is proposed to semantically align the vision and text prompts based on CLIP model, and reinforce each other. In addition, we propose an open-set knowledge transfer to transfer the category-specific discrepancies into the backbone network using knowledge distillation mechanism. A variety of quantitative and qualitative experiments show that our PLEor achieves promising performance on open-set fine-grained retrieval datasets.
****
Sibling-Attack: Rethinking Transferable Adversarial Attacks Against Face Recognition
Zexin Li, Bangjie Yin, Taiping Yao, Junfeng Guo, Shouhong Ding, Simin Chen, Cong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24626-24637
A hard challenge in developing practical face recognition (FR) attacks is due to the black-box nature of the target FR model, i.e., inaccessible gradient and parameter information to attackers. While recent research took an important step towards attacking black-box FR models through leveraging transferability, their performance is still limited, especially against online commercial FR systems tha

t can be pessimistic (e.g., a less than 50% ASR--attack success rate on average). Motivated by this, we present Sibling-Attack, a new FR attack technique for the first time explores a novel multi-task perspective (i.e., leveraging extra information from multi-correlated tasks to boost attacking transferability). Intuitively, Sibling-Attack selects a set of tasks correlated with FR and picks the Attribute Recognition (AR) task as the task used in Sibling-Attack based on theoretical and quantitative analysis. Sibling-Attack then develops an optimization framework that fuses adversarial gradient information through (1) constraining the cross-task features to be under the same space, (2) a joint-task meta optimization framework that enhances the gradient compatibility among tasks, and (3) a cross-task gradient stabilization method which mitigates the oscillation effect during attacking. Extensive experiments demonstrate that Sibling-Attack outperforms state-of-the-art FR attack techniques by a non-trivial margin, boosting ASR by 12.61% and 55.77% on average on state-of-the-art pre-trained FR models and two well-known, widely used commercial FR systems.
****
Enhanced Stable View Synthesis
Nishant Jain, Suryansh Kumar, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13208-13217
We introduce an approach to enhance the novel view synthesis from images taken from a freely moving camera. The introduced approach focuses on outdoor scenes where recovering accurate geometric scaffold and camera pose is challenging, leading to inferior results using the state-of-the-art stable view synthesis (SVS) method. SVS and related methods fail for outdoor scenes primarily due to (i) over-relying on the multiview stereo (MVS) for geometric scaffold recovery and (ii) assuming COLMAP computed camera poses as the best possible estimates, despite it being well-studied that MVS 3D reconstruction accuracy is limited to scene disparity and camera-pose accuracy is sensitive to key-point correspondence selection. This work proposes a principled way to enhance novel view synthesis solutions drawing inspiration from the basics of multiple view geometry. By leveraging the complementary behavior of MVS and monocular depth, we arrive at a better scene depth per view for nearby and far points, respectively. Moreover, our approach jointly refines camera poses with image-based rendering via multiple rotation averaging graph optimization. The recovered scene depth and the camera-pose help better view-dependent on-surface feature aggregation of the entire scene. Extensive evaluation of our approach on the popular benchmark dataset, such as Tanks and Temples, shows substantial improvement in view synthesis results compared to the prior art. For instance, our method shows 1.5 dB of PSNR improvement on the Tank and Temples. Similar statistics are observed when tested on other benchmark datasets such as FVS, Mip-NeRF 360, and DTU.
****
Breaching FedMD: Image Recovery via Paired-Logits Inversion Attack
Hideaki Takahashi, Jingjing Liu, Yang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12198-12207
Federated Learning with Model Distillation (FedMD) is a nascent collaborative learning paradigm, where only output logits of public datasets are transmitted as distilled knowledge, instead of passing on private model parameters that are susceptible to gradient inversion attacks, a known privacy risk in federated learning. In this paper, we found that even though sharing output logits of public datasets is safer than directly sharing gradients, there still exists a substantial risk of data exposure caused by carefully designed malicious attacks. Our study shows that a malicious server can inject a PLI (Paired-Logits Inversion) attack against FedMD and its variants by training an inversion neural network that exploits the confidence gap between the server and client models. Experiments on multiple facial recognition datasets validate that under FedMD-like schemes, by using paired server-client logits of public datasets only, the malicious server is able to reconstruct private images on all tested benchmarks with a high success rate.
****
TempSAL - Uncovering Temporal Information for Deep Saliency Prediction

Bahar Aydemir, Ludo Hoffstetter, Tong Zhang, Mathieu Salzmann, Sabine Süsstrunk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6461-6470
Deep saliency prediction algorithms complement the object recognition features, they typically rely on additional information such as scene context, semantic relationships, gaze direction, and object dissimilarity. However, none of these models consider the temporal nature of gaze shifts during image observation. We introduce a novel saliency prediction model that learns to output saliency maps in sequential time intervals by exploiting human temporal attention patterns. Our approach locally modulates the saliency predictions by combining the learned temporal maps. Our experiments show that our method outperforms the state-of-the-art models, including a multi-duration saliency model, on the SALICON benchmark and CodeCharts1k dataset. Our code is publicly available on GitHub.
****
Biomechanics-Guided Facial Action Unit Detection Through Force Modeling
Zijun Cui, Chenyi Kuang, Tian Gao, Kartik Talamadupula, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8694-8703
Existing AU detection algorithms are mainly based on appearance information extracted from 2D images, and well-established facial biomechanics that governs 3D facial skin deformation is rarely considered. In this paper, we propose a biomechanics-guided AU detection approach, where facial muscle activation forces are modelled, and are employed to predict AU activation. Specifically, our model consists of two branches: 3D physics branch and 2D image branch. In 3D physics branch, we first derive the Euler-Lagrange equation governing facial deformation. The Euler-Lagrange equation represented as an ordinary differential equation (ODE) is embedded into a differentiable ODE solver. Muscle activation forces together with other physics parameters are firstly regressed, and then are utilized to simulate 3D deformation by solving the ODE. By leveraging facial biomechanics, we obtain physically plausible facial muscle activation forces. 2D image branch compensates 3D physics branch by employing additional appearance information from 2D images. Both estimated forces and appearance features are employed for AU detection. The proposed approach achieves competitive AU detection performance on two benchmark datasets. Furthermore, by leveraging biomechanics, our approach achieves outstanding performance with reduced training data.
****
Equiangular Basis Vectors
Yang Shen, Xuhao Sun, Xiu-Shen Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11755-11765
We propose Equiangular Basis Vectors (EBVs) for classification tasks. In deep neural networks, models usually end with a k-way fully connected layer with softmax to handle different classification tasks. The learning objective of these methods can be summarized as mapping the learned feature representations to the samples' label space. While in metric learning approaches, the main objective is to learn a transformation function that maps training data points from the original space to a new space where similar points are closer while dissimilar points become farther apart. Different from previous methods, our EBVs generate normalized vector embeddings as "predefined classifiers" which are required to not only be with the equal status between each other, but also be as orthogonal as possible. By minimizing the spherical distance of the embedding of an input between its categorical EBV in training, the predictions can be obtained by identifying the categorical EBV with the smallest distance during inference. Various experiments on the ImageNet-1K dataset and other downstream tasks demonstrate that our method outperforms the general fully connected classifier while it does not introduce huge additional computation compared with classical metric learning methods. Our EBVs won the first place in the 2022 DIGIX Global AI Challenge, and our code is open-source and available at https://github.com/NJUST-VIPGroup/Equiangular-Basis-Vectors.
****
PIRLNav: Pretraining With Imitation and RL Finetuning for ObjectNav

Ram Ramrakhya, Dhruv Batra, Erik Wijmans, Abhishek Das; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17896-17906

We study ObjectGoal Navigation -- where a virtual robot situated in a new environment is asked to navigate to an object. Prior work has shown that imitation learning (IL) using behavior cloning (BC) on a dataset of human demonstrations achieves promising results. However, this has limitations -- 1) BC policies generalize poorly to new states, since the training mimics actions not their consequences, and 2) collecting demonstrations is expensive. On the other hand, reinforcement learning (RL) is trivially scalable, but requires careful reward engineering to achieve desirable behavior. We present PIRLNav, a two-stage learning scheme for BC pretraining on human demonstrations followed by RL-finetuning. This leads to a policy that achieves a success rate of 65.0% on ObjectNav (+5.0% absolute over previous state-of-the-art). Using this BC->RL training recipe, we present a rigorous empirical analysis of design choices. First, we investigate whether human demonstrations can be replaced with 'free' (automatically generated) sources of demonstrations, e.g. shortest paths (SP) or task-agnostic frontier exploration (FE) trajectories. We find that BC->RL on human demonstrations outperforms BC->RL on SP and FE trajectories, even when controlled for the same BC-pretraining success on train, and even on a subset of val episodes where BC-pretraining success favors the SP or FE policies. Next, we study how RL-finetuning performance scales with the size of the BC pretraining dataset. We find that as we increase the size of the BC-pretraining dataset and get to high BC accuracies, the improvements from RL-finetuning are smaller, and that 90% of the performance of our best BC->RL policy can be achieved with less than half the number of BC demonstrations. Finally, we analyze failure modes of our ObjectNav policies, and present guidelines for further improving them.

****

Megahertz Light Steering Without Moving Parts

Adithya Pediredla, Srinivasa G. Narasimhan, Maysamreza Chamanzar, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1-12

We introduce a light steering technology that operates at megahertz frequencies, has no moving parts, and costs less than a hundred dollars. Our technology can benefit many projector and imaging systems that critically rely on high-speed, reliable, low-cost, and wavelength-independent light steering, including laser scanning projectors, LiDAR sensors, and fluorescence microscopes. Our technology uses ultrasound waves to generate a spatiotemporally-varying refractive index field inside a compressible medium, such as water, turning the medium into a dynamic traveling lens. By controlling the electrical input of the ultrasound transducers that generate the waves, we can change the lens, and thus steer light, at the speed of sound (1.5 km/s in water). We build a physical prototype of this technology, use it to realize different scanning techniques at megahertz rates (three orders of magnitude faster than commercial alternatives such as galvo mirror scanners), and demonstrate proof-of-concept projector and LiDAR applications. To encourage further innovation towards this new technology, we derive the theory for its fundamental limits and develop a physically-accurate simulator for virtual design. Our technology offers a promising solution for achieving high-speed and low-cost light steering in a variety of applications.

****

Iterative Proposal Refinement for Weakly-Supervised Video Grounding

Meng Cao, Fangyun Wei, Can Xu, Xiubo Geng, Long Chen, Can Zhang, Yuexian Zou, Tao Shen, Daxin Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6524-6534

Weakly-Supervised Video Grounding (WSVG) aims to localize events of interest in untrimmed videos with only video-level annotations. To date, most of the state-of-the-art WSVG methods follow a two-stage pipeline, i.e., firstly generating potential temporal proposals and then grounding with these proposal candidates. Despite the recent progress, existing proposal generation methods suffer from two drawbacks: 1) lack of explicit correspondence modeling; and 2) partial coverage o

f complex events. To this end, we propose a novel IteRative prOposal refiNement network (dubbed as IRON) to gradually distill the prior knowledge into each proposal and encourage proposals with more complete coverage. Specifically, we set up two lightweight distillation branches to uncover the cross-modal correspondence on both the semantic and conceptual levels. Then, an iterative Label Propagation (LP) strategy is devised to prevent the network from focusing excessively on the most discriminative events instead of the whole sentence content. Precisely, during each iteration, the proposal with the minimal distillation loss and its adjacent ones are regarded as the positive samples, which refines proposal confidence scores in a cascaded manner. Extensive experiments and ablation studies on two challenging WSVG datasets have attested to the effectiveness of our IRON.
****
SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy
Jiafeng Li, Ying Wen, Lianghua He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6153-6162
Convolutional Neural Networks (CNNs) have achieved remarkable performance in various computer vision tasks but this comes at the cost of tremendous computational resources, partly due to convolutional layers extracting redundant features. Recent works either compress well-trained large-scale models or explore well-designed lightweight models. In this paper, we make an attempt to exploit spatial and channel redundancy among features for CNN compression and propose an efficient convolution module, called SCConv (Spatial and Channel reconstruction Convolution), to decrease redundant computing and facilitate representative feature learning. The proposed SCConv consists of two units: spatial reconstruction unit (SRU) and channel reconstruction unit (CRU). SRU utilizes a separate-and-reconstruct method to suppress the spatial redundancy while CRU uses a split-transform-and-fuse strategy to diminish the channel redundancy. In addition, SCConv is a plug-and-play architectural unit that can be used to replace standard convolution in various convolutional neural networks directly. Experimental results show that SCConv-embedded models are able to achieve better performance by reducing redundant features with significantly lower complexity and computational costs.
****
StyleGene: Crossover and Mutation of Region-Level Facial Genes for Kinship Face Synthesis
Hao Li, Xianxu Hou, Zepeng Huang, Linlin Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20960-20969
High-fidelity kinship face synthesis has many potential applications, such as kinship verification, missing child identification, and social media analysis. However, it is challenging to synthesize high-quality descendant faces with genetic relations due to the lack of large-scale, high-quality annotated kinship data. This paper proposes RFG (Region-level Facial Gene) extraction framework to address this issue. We propose to use IGE (Image-based Gene Encoder), LGE (Latent-based Gene Encoder) and Gene Decoder to learn the RFGs of a given face image, and the relationships between RFGs and the latent space of StyleGAN2. As cycle-like losses are designed to measure the L_2 distances between the output of Gene Decoder and image encoder, and that between the output of LGE and IGE, only face images are required to train our framework, i.e. no paired kinship face data is required. Based upon the proposed RFGs, a crossover and mutation module is further designed to inherit the facial parts of parents. A Gene Pool has also been used to introduce the variations into the mutation of RFGs. The diversity of the faces of descendants can thus be significantly increased. Qualitative, quantitative, and subjective experiments on FIW, TSKinFace, and FF-Databases clearly show that the quality and diversity of kinship faces generated by our approach are much better than the existing state-of-the-art methods.
****
Clothed Human Performance Capture With a Double-Layer Neural Radiance Fields
Kangkan Wang, Guofeng Zhang, Suxu Cong, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21098-21107
This paper addresses the challenge of capturing performance for the clothed huma

ns from sparse-view or monocular videos. Previous methods capture the performanc e of full humans with a personalized template or recover the garments from a sin gle frame with static human poses. However, it is inconvenient to extract cloth semantics and capture clothing motion with one-piece template, while single fram e-based methods may suffer from instable tracking across videos. To address thes e problems, we propose a novel method for human performance capture by tracking clothing and human body motion separately with a double-layer neural radiance fi elds (NeRFs). Specifically, we propose a double-layer NeRFs for the body and gar ments, and track the densely deforming template of the clothing and body by join tly optimizing the deformation fields and the canonical double-layer NeRFs. In t he optimization, we introduce a physics-aware cloth simulation network which can help generate physically plausible cloth dynamics and body-cloth interactions. Compared with existing methods, our method is fully differentiable and can captu re both the body and clothing motion robustly from dynamic videos. Also, our met hod represents the clothing with an independent NeRFs, allowing us to model impl icit fields of general clothes feasibly. The experimental evaluations validate i ts effectiveness on real multi-view or monocular videos.
****
NeuFace: Realistic 3D Neural Face Rendering From Multi-View Images
Mingwu Zheng, Haiyu Zhang, Hongyu Yang, Di Huang; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16868-1687 7
Realistic face rendering from multi-view images is beneficial to various compute r vision and graphics applications. Due to the complex spatially-varying reflect ance properties and geometry characteristics of faces, however, it remains chall enging to recover 3D facial representations both faithfully and efficiently in t he current studies. This paper presents a novel 3D face rendering model, namely NeuFace, to learn accurate and physically-meaningful underlying 3D representatio ns by neural rendering techniques. It naturally incorporates the neural BRDFs in to physically based rendering, capturing sophisticated facial geometry and appea rance clues in a collaborative manner. Specifically, we introduce an approximate d BRDF integration and a simple yet new low-rank prior, which effectively lower the ambiguities and boost the performance of the facial BRDFs. Extensive experim ents demonstrate the superiority of NeuFace in human face rendering, along with a decent generalization ability to common objects. Code is released at https://g ithub.com/aejion/NeuFace.
****
Cross-Guided Optimization of Radiance Fields With Multi-View Image Super-Resolut ion for High-Resolution Novel View Synthesis
Youngho Yoon, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2023, pp. 12428-12438
Novel View Synthesis (NVS) aims at synthesizing an image from an arbitrary viewp oint using multi-view images and camera poses. Among the methods for NVS, Neural Radiance Fields (NeRF) is capable of NVS for an arbitrary resolution as it lear ns a continuous volumetric representation. However, radiance fields rely heavily on the spectral characteristics of coordinate-based networks. Thus, there is a limit to improving the performance of high-resolution novel view synthesis (HRNV S). To solve this problem, we propose a novel framework using cross-guided optim ization of the single-image super-resolution (SISR) and radiance fields. We perf orm multi-view image super-resolution (MVSR) on train-view images during the rad iance fields optimization process. It derives the updated SR result by fusing th e feature map obtained from SISR and voxel-based uncertainty fields generated by integrated errors of train-view images. By repeating the updates during radianc e fields optimization, train-view images for radiance fields optimization have m ulti-view consistency and high-frequency details simultaneously, ultimately impr oving the performance of HRNVS. Experiments of HRNVS and MVSR on various benchma rk datasets show that the proposed method significantly surpasses existing metho ds.
****
Unified Pose Sequence Modeling

Lin Geng Foo, Tianjiao Li, Hossein Rahmani, Qiuhong Ke, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13019-13030
We propose a Unified Pose Sequence Modeling approach to unify heterogeneous human behavior understanding tasks based on pose data, e.g., action recognition, 3D pose estimation and 3D early action prediction. A major obstacle is that different pose-based tasks require different output data formats. Specifically, the action recognition and prediction tasks require class predictions as outputs, while 3D pose estimation requires a human pose output, which limits existing methods to leverage task-specific network architectures for each task. Hence, in this paper, we propose a novel Unified Pose Sequence (UPS) model to unify heterogeneous output formats for the aforementioned tasks by considering text-based action labels and coordinate-based human poses as language sequences. Then, by optimizing a single auto-regressive transformer, we can obtain a unified output sequence that can handle all the aforementioned tasks. Moreover, to avoid the interference brought by the heterogeneity between different tasks, a dynamic routing mechanism is also proposed to empower our UPS with the ability to learn which subsets of parameters should be shared among different tasks. To evaluate the efficacy of the proposed UPS, extensive experiments are conducted on four different tasks with four popular behavior understanding benchmarks.
****
Probability-Based Global Cross-Modal Upsampling for Pansharpening
Zeyu Zhu, Xiangyong Cao, Man Zhou, Junhao Huang, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14039-14048
Pansharpening is an essential preprocessing step for remote sensing image processing. Although deep learning (DL) approaches performed well on this task, current upsampling methods used in these approaches only utilize the local information of each pixel in the low-resolution multispectral (LRMS) image while neglecting to exploit its global information as well as the cross-modal information of the guiding panchromatic (PAN) image, which limits their performance improvement. To address this issue, this paper develops a novel probability-based global cross-modal upsampling (PGCU) method for pan-sharpening. Precisely, we first formulate the PGCU method from a probabilistic perspective and then design an efficient network module to implement it by fully utilizing the information mentioned above while simultaneously considering the channel specificity. The PGCU module consists of three blocks, i.e., information extraction (IE), distribution and expectation estimation (DEE), and fine adjustment (FA). Extensive experiments verify the superiority of the PGCU method compared with other popular upsampling methods. Additionally, experiments also show that the PGCU module can help improve the performance of existing SOTA deep learning pansharpening methods. The codes are available at https://github.com/Zeyu-Zhu/PGCU.
****
Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation
Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6914-6924
The CLIP model has been recently proven to be very effective for a variety of cross-modal tasks, including the evaluation of captions generated from vision-and-language architectures. In this paper, we propose a new recipe for a contrastive-based evaluation metric for image captioning, namely Positive-Augmented Contrastive learning Score (PAC-S), that in a novel way unifies the learning of a contrastive visual-semantic space with the addition of generated images and text on curated data. Experiments spanning several datasets demonstrate that our new metric achieves the highest correlation with human judgments on both images and videos, outperforming existing reference-based metrics like CIDEr and SPICE and reference-free metrics like CLIP-Score. Finally, we test the system-level correlation of the proposed metric when considering popular image captioning approaches, and assess the impact of employing different cross-modal features. Our source cod

e and trained models are publicly available at: https://github.com/aimagelab/pac score.
****
Rethinking Domain Generalization for Face Anti-Spoofing: Separability and Alignment
Yiyou Sun, Yaojie Liu, Xiaoming Liu, Yixuan Li, Wen-Sheng Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24563-24574
This work studies the generalization issue of face anti-spoofing (FAS) models on domain gaps, such as image resolution, blurriness and sensor variations. Most prior works regard domain-specific signals as a negative impact, and apply metric learning or adversarial losses to remove it from feature representation. Though learning a domain-invariant feature space is viable for the training data, we show that the feature shift still exists in an unseen test domain, which backfires on the generalizability of the classifier. In this work, instead of constructing a domain-invariant feature space, we encourage domain separability while aligning the live-to-spoof transition (i.e., the trajectory from live to spoof) to be the same for all domains. We formulate this FAS strategy of separability and alignment (SA-FAS) as a problem of invariant risk minimization (IRM), and learn domain-variant feature representation but domain-invariant classifier. We demonstrate the effectiveness of SA-FAS on challenging cross-domain FAS datasets and establish state-of-the-art performance.
****
SMOC-Net: Leveraging Camera Pose for Self-Supervised Monocular Object Pose Estimation
Tao Tan, Qiulei Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21307-21316
Recently, self-supervised 6D object pose estimation, where synthetic images with object poses (sometimes jointly with un-annotated real images) are used for training, has attracted much attention in computer vision. Some typical works in literature employ a time-consuming differentiable renderer for object pose prediction at the training stage, so that (i) their performances on real images are generally limited due to the gap between their rendered images and real images and (ii) their training process is computationally expensive. To address the two problems, we propose a novel Network for Self-supervised Monocular Object pose estimation by utilizing the predicted Camera poses from un-annotated real images, called SMOC-Net. The proposed network is explored under a knowledge distillation framework, consisting of a teacher model and a student model. The teacher model contains a backbone estimation module for initial object pose estimation, and an object pose refiner for refining the initial object poses using a geometric constraint (called relative-pose constraint) derived from relative camera poses. The student model gains knowledge for object pose estimation from the teacher model by imposing the relative-pose constraint. Thanks to the relative-pose constraint, SMOC-Net could not only narrow the domain gap between synthetic and real data but also reduce the training cost. Experimental results on two public datasets demonstrate that SMOC-Net outperforms several state-of-the-art methods by a large margin while requiring much less training time than the differentiable-renderer-based methods.
****
FAC: 3D Representation Learning via Foreground Aware Feature Contrast
Kangcheng Liu, Aoran Xiao, Xiaoqin Zhang, Shijian Lu, Ling Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9476-9485
Contrastive learning has recently demonstrated great potential for unsupervised pre-training in 3D scene understanding tasks. However, most existing work randomly selects point features as anchors while building contrast, leading to a clear bias toward background points that often dominate in 3D scenes. Also, object awareness and foreground-to-background discrimination are neglected, making contrastive learning less effective. To tackle these issues, we propose a general foreground-aware feature contrast (FAC) framework to learn more effective point clou

d representations in pre-training. FAC consists of two novel contrast designs to construct more effective and informative contrast pairs. The first is building positive pairs within the same foreground segment where points tend to have the same semantics. The second is that we prevent over-discrimination between 3D segments/objects and encourage foreground-to-background distinctions at the segment level with adaptive feature learning in a Siamese correspondence network, which adaptively learns feature correlations within and across point cloud views effectively. Visualization with point activation maps shows that our contrast pairs capture clear correspondences among foreground regions during pre-training. Quantitative experiments also show that FAC achieves superior knowledge transfer and data efficiency in various downstream 3D semantic segmentation and object detection tasks. All codes, data, and models are available at:https://github.com/KangchengLiu/FAC_Foreground_Aware_Contrast.
****
Improving Visual Representation Learning Through Perceptual Understanding
Samyakh Tukra, Frederick Hoffman, Ken Chatfield; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14486-14495
We present an extension to masked autoencoders (MAE) which improves on the representations learnt by the model by explicitly encouraging the learning of higher scene-level features. We do this by: (i) the introduction of a perceptual similarity term between generated and real images (ii) incorporating several techniques from the adversarial training literature including multi-scale training and adaptive discriminator augmentation. The combination of these results in not only better pixel reconstruction but also representations which appear to capture better higher-level details within images. More consequentially, we show how our method, Perceptual MAE, leads to better performance when used for downstream tasks outperforming previous methods. We achieve 78.1% top-1 accuracy linear probing on ImageNet-1K and up to 88.1% when fine-tuning, with similar results for other downstream tasks, all without use of additional pre-trained models or data.
****
3D Cinemagraphy From a Single Image
Xingyi Li, Zhiguo Cao, Huiqiang Sun, Jianming Zhang, Ke Xian, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4595-4605
We present 3D Cinemagraphy, a new technique that marries 2D image animation with 3D photography. Given a single still image as input, our goal is to generate a video that contains both visual content animation and camera motion. We empirically find that naively combining existing 2D image animation and 3D photography methods leads to obvious artifacts or inconsistent animation. Our key insight is that representing and animating the scene in 3D space offers a natural solution to this task. To this end, we first convert the input image into feature-based layered depth images using predicted depth values, followed by unprojecting them to a feature point cloud. To animate the scene, we perform motion estimation and lift the 2D motion into the 3D scene flow. Finally, to resolve the problem of hole emergence as points move forward, we propose to bidirectionally displace the point cloud as per the scene flow and synthesize novel views by separately projecting them into target image planes and blending the results. Extensive experiments demonstrate the effectiveness of our method. A user study is also conducted to validate the compelling rendering results of our method.
****
Learning Bottleneck Concepts in Image Classification
Bowen Wang, Liangzhi Li, Yuta Nakashima, Hajime Nagahara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10962-10971
Interpreting and explaining the behavior of deep neural networks is critical for many tasks. Explainable AI provides a way to address this challenge, mostly by providing per-pixel relevance to the decision. Yet, interpreting such explanations may require expert knowledge. Some recent attempts toward interpretability adopt a concept-based framework, giving a higher-level relationship between some concepts and model decisions. This paper proposes Bottleneck Concept Learner (Bot

CL), which represents an image solely by the presence/absence of concepts learne d through training over the target task without explicit supervision over the co ncepts. It uses self-supervision and tailored regularizers so that learned conce pts can be human-understandable. Using some image classification tasks as our te stbed, we demonstrate BotCL's potential to rebuild neural networks for better in terpretability.
****
Inversion-Based Style Transfer With Diffusion Models
Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, Ch angsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2023, pp. 10146-10156
The artistic style within a painting is the means of expression, which includes not only the painting material, colors, and brushstrokes, but also the high-leve l attributes, including semantic elements and object shapes. Previous arbitrary example-guided artistic image generation methods often fail to control shape cha nges or convey elements. Pre-trained text-to-image synthesis diffusion probabili stic models have achieved remarkable quality but often require extensive textual descriptions to accurately portray the attributes of a particular painting.The uniqueness of an artwork lies in the fact that it cannot be adequately explained with normal language. Our key idea is to learn the artistic style directly from a single painting and then guide the synthesis without providing complex textua l descriptions. Specifically, we perceive style as a learnable textual descripti on of a painting.We propose an inversion-based style transfer method (InST), whi ch can efficiently and accurately learn the key information of an image, thus ca pturing and transferring the artistic style of a painting. We demonstrate the qu ality and efficiency of our method on numerous paintings of various artists and styles. Codes are available at https://github.com/zyxElsa/InST.
****
Learning Human Mesh Recovery in 3D Scenes
Zehong Shen, Zhi Cen, Sida Peng, Qing Shuai, Hujun Bao, Xiaowei Zhou; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17038-17047
We present a novel method for recovering the absolute pose and shape of a human in a pre-scanned scene given a single image. Unlike previous methods that perfor m sceneaware mesh optimization, we propose to first estimate absolute position a nd dense scene contacts with a sparse 3D CNN, and later enhance a pretrained hum an mesh recovery network by cross-attention with the derived 3D scene cues. Join t learning on images and scene geometry enables our method to reduce the ambigui ty caused by depth and occlusion, resulting in more reasonable global postures a nd contacts. Encoding scene-aware cues in the network also allows the proposed m ethod to be optimization-free, and opens up the opportunity for real-time applic ations. The experiments show that the proposed network is capable of recovering accurate and physically-plausible meshes by a single forward pass and outperform s state-of-the-art methods in terms of both accuracy and speed. Code is availabl e on our project page: https://zju3dv.github.io/sahmr/.
****
Learning Locally Editable Virtual Humans
Hsuan-I Ho, Lixin Xue, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21024-21035
In this paper, we propose a novel hybrid representation and end-to-end trainable network architecture to model fully editable and customizable neural avatars. A t the core of our work lies a representation that combines the modeling power of neural fields with the ease of use and inherent 3D consistency of skinned meshe s. To this end, we construct a trainable feature codebook to store local geometr y and texture features on the vertices of a deformable body model, thus exploiti ng its consistent topology under articulation. This representation is then emplo yed in a generative auto-decoder architecture that admits fitting to unseen scan s and sampling of realistic avatars with varied appearances and geometries. Furt hermore, our representation allows local editing by swapping local features betw een 3D assets. To verify our method for avatar creation and editing, we contribu

te a new high-quality dataset, dubbed CustomHumans, for training and evaluation. Our experiments quantitatively and qualitatively show that our method generates diverse detailed avatars and achieves better model fitting performance compared to state-of-the-art methods. Our code and dataset are available at https://ait.ethz.ch/custom-humans.
****
Learning Imbalanced Data With Vision Transformers
Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, Chun Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15793-15803
The real-world data tends to be heavily imbalanced and severely skew the data-driven deep neural networks, which makes Long-Tailed Recognition (LTR) a massive challenging task. Existing LTR methods seldom train Vision Transformers (ViTs) with Long-Tailed (LT) data, while the off-the-shelf pretrain weight of ViTs always leads to unfair comparisons. In this paper, we systematically investigate the ViTs' performance in LTR and propose LiVT to train ViTs from scratch only with LT data. With the observation that ViTs suffer more severe LTR problems, we conduct Masked Generative Pretraining (MGP) to learn generalized features. With ample and solid evidence, we show that MGP is more robust than supervised manners. Although Binary Cross Entropy (BCE) loss performs well with ViTs, it struggles on the LTR tasks. We further propose the balanced BCE to ameliorate it with strong theoretical groundings. Specially, we derive the unbiased extension of Sigmoid and compensate extra logit margins for deploying it. Our Bal-BCE contributes to the quick convergence of ViTs in just a few epochs. Extensive experiments demonstrate that with MGP and Bal-BCE, LiVT successfully trains ViTs well without any additional data and outperforms comparable state-of-the-art methods significantly, e.g., our ViT-B achieves 81.0% Top-1 accuracy in iNaturalist 2018 without bells and whistles. Code is available at https://github.com/XuZhengzhuo/LiVT.
****
AttriCLIP: A Non-Incremental Learner for Incremental Knowledge Learning
Runqi Wang, Xiaoyue Duan, Guoliang Kang, Jianzhuang Liu, Shaohui Lin, Songcen Xu, Jinhu Lü, Baochang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3654-3663
Continual learning aims to enable a model to incrementally learn knowledge from sequentially arrived data. Previous works adopt the conventional classification architecture, which consists of a feature extractor and a classifier. The feature extractor is shared across sequentially arrived tasks or classes, but one specific group of weights of the classifier corresponding to one new class should be incrementally expanded. Consequently, the parameters of a continual learner gradually increase. Moreover, as the classifier contains all historical arrived classes, a certain size of the memory is usually required to store rehearsal data to mitigate classifier bias and catastrophic forgetting. In this paper, we propose a non-incremental learner, named AttriCLIP, to incrementally extract knowledge of new classes or tasks. Specifically, AttriCLIP is built upon the pre-trained visual-language model CLIP. Its image encoder and text encoder are fixed to extract features from both images and text prompts. Each text prompt consists of a category name and a fixed number of learnable parameters which are selected from our designed attribute bank and serve as attributes. As we compute the visual and textual similarity for classification, AttriCLIP is a non-incremental learner. The attribute prompts, which encode the common knowledge useful for classification, can effectively mitigate the catastrophic forgetting and avoid constructing a replay memory. We empirically evaluate our AttriCLIP and compare it with CLIP-based and previous state-of-the-art continual learning methods in realistic settings with domain-shift and long-sequence learning. The results show that our method performs favorably against previous state-of-the-arts.
****
PHA: Patch-Wise High-Frequency Augmentation for Transformer-Based Person Re-Identification
Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, Shiliang Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p

p. 14133-14142
Although recent studies empirically show that injecting Convolutional Neural Net works (CNNs) into Vision Transformers (ViTs) can improve the performance of pers on re-identification, the rationale behind it remains elusive. From a frequency perspective, we reveal that ViTs perform worse than CNNs in preserving key high-frequency components (e.g, clothes texture details) since high-frequency compone nts are inevitably diluted by low-frequency ones due to the intrinsic Self-Atten tion within ViTs. To remedy such inadequacy of the ViT, we propose a Patch-wise High-frequency Augmentation (PHA) method with two core designs. First, to enhanc e the feature representation ability of high-frequency components, we split patc hes with high-frequency components by the Discrete Haar Wavelet Transform, then empower the ViT to take the split patches as auxiliary input. Second, to prevent high-frequency components from being diluted by low-frequency ones when taking the entire sequence as input during network optimization, we propose a novel pat ch-wise contrastive loss. From the view of gradient optimization, it acts as an implicit augmentation to improve the representation ability of key high-frequenc y components. This benefits the ViT to capture key high-frequency components to extract discriminative person representations. PHA is necessary during training and can be removed during inference, without bringing extra complexity. Extensiv e experiments on widely-used ReID datasets validate the effectiveness of our met hod.
****
StyleRes: Transforming the Residuals for Real Image Editing With StyleGAN
Hamza Pehlivan, Yusuf Dalva, Aysegul Dundar; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1828-1837
We present a novel image inversion framework and a training pipeline to achieve high-fidelity image inversion with high-quality attribute editing. Inverting rea l images into StyleGAN's latent space is an extensively studied problem, yet the trade-off between the image reconstruction fidelity and image editing quality r emains an open challenge. The low-rate latent spaces are limited in their expres siveness power for high-fidelity reconstruction. On the other hand, high-rate la tent spaces result in degradation in editing quality. In this work, to achieve h igh-fidelity inversion, we learn residual features in higher latent codes that l ower latent codes were not able to encode. This enables preserving image details in reconstruction. To achieve high-quality editing, we learn how to transform t he residual features for adapting to manipulations in latent codes. We train the framework to extract residual features and transform them via a novel architect ure pipeline and cycle consistency losses. We run extensive experiments and comp are our method with state-of-the-art inversion methods. Qualitative metrics and visual comparisons show significant improvements.
****
Diffusion Video Autoencoders: Toward Temporally Consistent Face Video Editing vi a Disentangled Video Encoding
Gyeongman Kim, Hajin Shim, Hyunsu Kim, Yunjey Choi, Junho Kim, Eunho Yang; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2023, pp. 6091-6100
Inspired by the impressive performance of recent face image editing methods, sev eral studies have been naturally proposed to extend these methods to the face vi deo editing task. One of the main challenges here is temporal consistency among edited frames, which is still unresolved. To this end, we propose a novel face v ideo editing framework based on diffusion autoencoders that can successfully ext ract the decomposed features - for the first time as a face video editing model - of identity and motion from a given video. This modeling allows us to edit the video by simply manipulating the temporally invariant feature to the desired di rection for the consistency. Another unique strength of our model is that, since our model is based on diffusion models, it can satisfy both reconstruction and edit capabilities at the same time, and is robust to corner cases in wild face v ideos (e.g. occluded faces) unlike the existing GAN-based methods.
****
Learning Instance-Level Representation for Large-Scale Multi-Modal Pretraining i

n E-Commerce
Yang Jin, Yongzhi Li, Zehuan Yuan, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11060-11069
This paper aims to establish a generic multi-modal foundation model that has the scalable capability to massive downstream applications in E-commerce. Recently, large-scale vision-language pretraining approaches have achieved remarkable advances in the general domain. However, due to the significant differences between natural and product images, directly applying these frameworks for modeling image-level representations to E-commerce will be inevitably sub-optimal. To this end, we propose an instance-centric multi-modal pretraining paradigm called ECLIP in this work. In detail, we craft a decoder architecture that introduces a set of learnable instance queries to explicitly aggregate instance-level semantics. Moreover, to enable the model to focus on the desired product instance without reliance on expensive manual annotations, two specially configured pretext tasks are further proposed. Pretrained on the 100 million E-commerce-related data, ECLIP successfully extracts more generic, semantic-rich, and robust representations. Extensive experimental results show that, without further fine-tuning, ECLIP surpasses existing methods by a large margin on a broad range of downstream tasks, demonstrating the strong transferability to real-world E-commerce applications.
****
Conditional Text Image Generation With Diffusion Models
Yuanzhi Zhu, Zhaohai Li, Tianwei Wang, Mengchao He, Cong Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14235-14245
Current text recognition systems, including those for handwritten scripts and scene text, have relied heavily on image synthesis and augmentation, since it is difficult to realize real-world complexity and diversity through collecting and annotating enough real text images. In this paper, we explore the problem of text image generation, by taking advantage of the powerful abilities of Diffusion Models in generating photo-realistic and diverse image samples with given conditions, and propose a method called Conditional Text Image Generation with Diffusion Models (CTIG-DM for short). To conform to the characteristics of text images, we devise three conditions: image condition, text condition, and style condition, which can be used to control the attributes, contents, and styles of the samples in the image generation process. Specifically, four text image generation modes, namely: (1) synthesis mode, (2) augmentation mode, (3) recovery mode, and (4) imitation mode, can be derived by combining and configuring these three conditions. Extensive experiments on both handwritten and scene text demonstrate that the proposed CTIG-DM is able to produce image samples that simulate real-world complexity and diversity, and thus can boost the performance of existing text recognizers. Besides, CTIG-DM shows its appealing potential in domain adaptation and generating images containing Out-Of-Vocabulary (OOV) words.
****
AnchorFormer: Point Cloud Completion From Discriminative Nodes
Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13581-13590
Point cloud completion aims to recover the completed 3D shape of an object from its partial observation. A common strategy is to encode the observed points to a global feature vector and then predict the complete points through a generative process on this vector. Nevertheless, the results may suffer from the high-quality shape generation problem due to the fact that a global feature vector cannot sufficiently characterize diverse patterns in one object. In this paper, we present a new shape completion architecture, namely AnchorFormer, that innovatively leverages pattern-aware discriminative nodes, i.e., anchors, to dynamically capture regional information of objects. Technically, AnchorFormer models the regional discrimination by learning a set of anchors based on the point features of the input partial observation. Such anchors are scattered to both observed and unobserved locations through estimating particular offsets, and form sparse points

together with the down-sampled points of the input observation. To reconstruct the fine-grained object patterns, AnchorFormer further employs a modulation scheme to morph a canonical 2D grid at individual locations of the sparse points into a detailed 3D structure. Extensive experiments on the PCN, ShapeNet-55/34 and KITTI datasets quantitatively and qualitatively demonstrate the efficacy of AnchorFormer over the state-of-the-art point cloud completion approaches. Source code is available at https://github.com/chenzhik/AnchorFormer.

****

Co-SLAM: Joint Coordinate and Sparse Parametric Encodings for Neural Real-Time SLAM

Hengyi Wang, Jingwen Wang, Lourdes Agapito; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13293-13302

We present Co-SLAM, a neural RGB-D SLAM system based on a hybrid representation, that performs robust camera tracking and high-fidelity surface reconstruction in real time. Co-SLAM represents the scene as a multi-resolution hash-grid to exploit its high convergence speed and ability to represent high-frequency local features. In addition, Co-SLAM incorporates one-blob encoding, to encourage surface coherence and completion in unobserved areas. This joint parametric-coordinate encoding enables real-time and robust performance by bringing the best of both worlds: fast convergence and surface hole filling. Moreover, our ray sampling strategy allows Co-SLAM to perform global bundle adjustment over all keyframes instead of requiring keyframe selection to maintain a small number of active keyframes as competing neural SLAM approaches do. Experimental results show that Co-SLAM runs at 10-17Hz and achieves state-of-the-art scene reconstruction results, and competitive tracking performance in various datasets and benchmarks (ScanNet, TUM, Replica, Synthetic RGBD). Project page: https://hengyiwang.github.io/projects/CoSLAM

****

SIM: Semantic-Aware Instance Mask Generation for Box-Supervised Instance Segmentation

Ruihuang Li, Chenhang He, Yabin Zhang, Shuai Li, Liyi Chen, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7193-7203

Weakly supervised instance segmentation using only bounding box annotations has recently attracted much research attention. Most of the current efforts leverage low-level image features as extra supervision without explicitly exploiting the high-level semantic information of the objects, which will become ineffective when the foreground objects have similar appearances to the background or other objects nearby. We propose a new box-supervised instance segmentation approach by developing a Semantic-aware Instance Mask (SIM) generation paradigm. Instead of heavily relying on local pair-wise affinities among neighboring pixels, we construct a group of category-wise feature centroids as prototypes to identify foreground objects and assign them semantic-level pseudo labels. Considering that the semantic-aware prototypes cannot distinguish different instances of the same semantics, we propose a self-correction mechanism to rectify the falsely activated regions while enhancing the correct ones. Furthermore, to handle the occlusions between objects, we tailor the Copy-Paste operation for the weakly-supervised instance segmentation task to augment challenging training data. Extensive experimental results demonstrate the superiority of our proposed SIM approach over other state-of-the-art methods. The source code: https://github.com/lslrh/SIM.

****

Compression-Aware Video Super-Resolution

Yingwei Wang, Takashi Isobe, Xu Jia, Xin Tao, Huchuan Lu, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2012-2021

Videos stored on mobile devices or delivered on the Internet are usually in compressed format and are of various unknown compression parameters, but most video super-resolution (VSR) methods often assume ideal inputs resulting in large performance gap between experimental settings and real-world applications. In spite of a few pioneering works being proposed recently to super-resolve the compresse

d videos, they are not specially designed to deal with videos of various levels of compression. In this paper, we propose a novel and practical compression-aware video super-resolution model, which could adapt its video enhancement process to the estimated compression level. A compression encoder is designed to model compression levels of input frames, and a base VSR model is then conditioned on the implicitly computed representation by inserting compression-aware modules. In addition, we propose to further strengthen the VSR model by taking full advantage of meta data that is embedded naturally in compressed video streams in the procedure of information fusion. Extensive experiments are conducted to demonstrate the effectiveness and efficiency of the proposed method on compressed VSR benchmarks.
****
PillarNeXt: Rethinking Network Designs for 3D Object Detection in LiDAR Point Clouds
Jinyu Li, Chenxu Luo, Xiaodong Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17567-17576
In order to deal with the sparse and unstructured raw point clouds, most LiDAR based 3D object detection research focuses on designing dedicated local point aggregators for fine-grained geometrical modeling. In this paper, we revisit the local point aggregators from the perspective of allocating computational resources. We find that the simplest pillar based models perform surprisingly well considering both accuracy and latency. Additionally, we show that minimal adaptions from the success of 2D object detection, such as enlarging receptive field, significantly boost the performance. Extensive experiments reveal that our pillar based networks with modernized designs in terms of architecture and training render the state-of-the-art performance on two popular benchmarks: Waymo Open Dataset and nuScenes. Our results challenge the common intuition that detailed geometry modeling is essential to achieve high performance for 3D object detection.
****
Regularization of Polynomial Networks for Image Recognition
Grigorios G. Chrysos, Bohan Wang, Jiankang Deng, Volkan Cevher; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16123-16132
Deep Neural Networks (DNNs) have obtained impressive performance across tasks, however they still remain as black boxes, e.g., hard to theoretically analyze. At the same time, Polynomial Networks (PNs) have emerged as an alternative method with a promising performance and improved interpretability but have yet to reach the performance of the powerful DNN baselines. In this work, we aim to close this performance gap. We introduce a class of PNs, which are able to reach the performance of ResNet across a range of six benchmarks. We demonstrate that strong regularization is critical and conduct an extensive study of the exact regularization schemes required to match performance. To further motivate the regularization schemes, we introduce D-PolyNets that achieve a higher-degree of expansion than previously proposed polynomial networks. D-PolyNets are more parameter-efficient while achieving a similar performance as other polynomial networks. We expect that our new models can lead to an understanding of the role of elementwise activation functions (which are no longer required for training PNs). The source code is available at https://github.com/grigorisg9gr/regularized_polynomials.
****
Incremental 3D Semantic Scene Graph Prediction From RGB Sequences
Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5064-5074
3D semantic scene graphs are a powerful holistic representation as they describe the individual objects and depict the relation between them. They are compact high-level graphs that enable many tasks requiring scene reasoning. In real-world settings, existing 3D estimation methods produce robust predictions that mostly rely on dense inputs. In this work, we propose a real-time framework that incrementally builds a consistent 3D semantic scene graph of a scene given an RGB image sequence. Our method consists of a novel incremental entity estimation pipeli

ne and a scene graph prediction network. The proposed pipeline simultaneously re constructs a sparse point map and fuses entity estimation from the input images. The proposed network estimates 3D semantic scene graphs with iterative message passing using multi-view and geometric features extracted from the scene entitie s. Extensive experiments on the 3RScan dataset show the effectiveness of the pro posed method in this challenging task, outperforming state-of-the-art approaches .
****
EfficientViT: Memory Efficient Vision Transformer With Cascaded Group Attention
Xinyu Liu, Houwen Peng, Ningxin Zheng, Yuqing Yang, Han Hu, Yixuan Yuan; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14420-14430
Vision transformers have shown great success due to their high model capabilitie s. However, their remarkable performance is accompanied by heavy computation cos ts, which makes them unsuitable for real-time applications. In this paper, we pr opose a family of high-speed vision transformers named EfficientViT. We find tha t the speed of existing transformer models is commonly bounded by memory ineffic ient operations, especially the tensor reshaping and element-wise functions in M HSA. Therefore, we design a new building block with a sandwich layout, i.e., usi ng a single memory-bound MHSA between efficient FFN layers, which improves memor y efficiency while enhancing channel communication. Moreover, we discover that t he attention maps share high similarities across heads, leading to computational redundancy. To address this, we present a cascaded group attention module feedi ng attention heads with different splits of the full feature, which not only sav es computation cost but also improves attention diversity. Comprehensive experim ents demonstrate EfficientViT outperforms existing efficient models, striking a good trade-off between speed and accuracy. For instance, our EfficientViT-M5 sur passes MobileNetV3-Large by 1.9% in accuracy, while getting 40.4% and 45.2% high er throughput on Nvidia V100 GPU and Intel Xeon CPU, respectively. Compared to t he recent efficient model MobileViT-XXS, EfficientViT-M2 achieves 1.8% superior accuracy, while running 5.8x/3.7x faster on the GPU/CPU, and 7.4x faster when co nverted to ONNX format. Code and models will be available soon.
****
VLPD: Context-Aware Pedestrian Detection via Vision-Language Semantic Self-Super vision
Mengyin Liu, Jie Jiang, Chao Zhu, Xu-Cheng Yin; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6662-6671
Detecting pedestrians accurately in urban scenes is significant for realistic ap plications like autonomous driving or video surveillance. However, confusing hum an-like objects often lead to wrong detections, and small scale or heavily occlu ded pedestrians are easily missed due to their unusual appearances. To address t hese challenges, only object regions are inadequate, thus how to fully utilize m ore explicit and semantic contexts becomes a key problem. Meanwhile, previous co ntext-aware pedestrian detectors either only learn latent contexts with visual c lues, or need laborious annotations to obtain explicit and semantic contexts. Th erefore, we propose in this paper a novel approach via Vision-Language semantic self-supervision for context-aware Pedestrian Detection (VLPD) to model explicit ly semantic contexts without any extra annotations. Firstly, we propose a self-s upervised Vision-Language Semantic (VLS) segmentation method, which learns both fully-supervised pedestrian detection and contextual segmentation via self-gener ated explicit labels of semantic classes by vision-language models. Furthermore, a self-supervised Prototypical Semantic Contrastive (PSC) learning method is pr oposed to better discriminate pedestrians and other classes, based on more expli cit and semantic contexts obtained from VLS. Extensive experiments on popular be nchmarks show that our proposed VLPD achieves superior performances over the pre vious state-of-the-arts, particularly under challenging circumstances like small scale and heavy occlusion. Code is available at https://github.com/lmy98129/VLP D.
****
TexPose: Neural Texture Learning for Self-Supervised 6D Object Pose Estimation

Hanzhi Chen, Fabian Manhardt, Nassir Navab, Benjamin Busam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4841-4852

In this paper, we introduce neural texture learning for 6D object pose estimation from synthetic data and a few unlabelled real images. Our major contribution is a novel learning scheme which removes the drawbacks of previous works, namely the strong dependency on co-modalities or additional refinement. These have been previously necessary to provide training signals for convergence. We formulate such a scheme as two sub-optimisation problems on texture learning and pose learning. We separately learn to predict realistic texture of objects from real image collections and learn pose estimation from pixel-perfect synthetic data. Combining these two capabilities allows then to synthesise photorealistic novel views to supervise the pose estimator with accurate geometry. To alleviate pose noise and segmentation imperfection present during the texture learning phase, we propose a surfel-based adversarial training loss together with texture regularisation from synthetic data. We demonstrate that the proposed approach significantly outperforms the recent state-of-the-art methods without ground-truth pose annotations and demonstrates substantial generalisation improvements towards unseen scenes. Remarkably, our scheme improves the adopted pose estimators substantially even when initialised with much inferior performance.

****

LINe: Out-of-Distribution Detection by Leveraging Important Neurons

Yong Hyun Ahn, Gyeong-Moon Park, Seong Tae Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19852-19862

It is important to quantify the uncertainty of input samples, especially in mission-critical domains such as autonomous driving and healthcare, where failure predictions on out-of-distribution (OOD) data are likely to cause big problems. OOD detection problem fundamentally begins in that the model cannot express what it is not aware of. Post-hoc OOD detection approaches are widely explored because they do not require an additional re-training process which might degrade the model's performance and increase the training cost. In this study, from the perspective of neurons in the deep layer of the model representing high-level features, we introduce a new aspect for analyzing the difference in model outputs between in-distribution data and OOD data. We propose a novel method, Leveraging Important Neurons (LINe), for post-hoc Out of distribution detection. Shapley value-based pruning reduces the effects of noisy outputs by selecting only high-contribution neurons for predicting specific classes of input data and masking the rest. Activation clipping fixes all values above a certain threshold into the same value, allowing LINe to treat all the class-specific features equally and just consider the difference between the number of activated feature differences between in-distribution and OOD data. Comprehensive experiments verify the effectiveness of the proposed method by outperforming state-of-the-art post-hoc OOD detection methods on CIFAR-10, CIFAR-100, and ImageNet datasets.

****

DynIBaR: Neural Dynamic Image-Based Rendering

Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, Noah Snavely; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4273-4284

We address the problem of synthesizing novel views from a monocular video depicting a complex dynamic scene. State-of-the-art methods based on temporally varying Neural Radiance Fields (aka dynamic NeRFs) have shown impressive results on this task. However, for long videos with complex object motions and uncontrolled camera trajectories,these methods can produce blurry or inaccurate renderings, hampering their use in real-world applications. Instead of encoding the entire dynamic scene within the weights of MLPs, we present a new approach that addresses these limitations by adopting a volumetric image-based rendering framework that synthesizes new viewpoints by aggregating features from nearby views in a scene motion-aware manner.Our system retains the advantages of prior methods in its ability to model complex scenes and view-dependent effects,but also enables synthesizing photo-realistic novel views from long videos featuring complex scene dyna

mics with unconstrained camera trajectories. We demonstrate significant improvements over state-of-the-art methods on dynamic scene datasets, and also apply our approach to in-the-wild videos with challenging camera and object motion, where prior methods fail to produce high-quality renderings
****
Unsupervised Object Localization: Observing the Background To Discover Objects
Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobecký, Éloi Zablocki, Patrick Pérez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3176-3186
Recent advances in self-supervised visual representation learning have paved the way for unsupervised methods tackling tasks such as object discovery and instance segmentation. However, discovering objects in an image with no supervision is a very hard task; what are the desired objects, when to separate them into parts, how many are there, and of what classes? The answers to these questions depend on the tasks and datasets of evaluation. In this work, we take a different approach and propose to look for the background instead. This way, the salient objects emerge as a by-product without any strong assumption on what an object should be. We propose FOUND, a simple model made of a single conv 1x1 initialized with coarse background masks extracted from self-supervised patch-based representations. After fast training and refining these seed masks, the model reaches state-of-the-art results on unsupervised saliency detection and object discovery benchmarks. Moreover, we show that our approach yields good results in the unsupervised semantic segmentation retrieval task. The code to reproduce our results is available at https://github.com/valeoai/FOUND.
****
Transforming Radiance Field With Lipschitz Network for Photorealistic 3D Scene Stylization
Zicheng Zhang, Yinglu Liu, Congying Han, Yingwei Pan, Tiande Guo, Ting Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20712-20721
Recent advances in 3D scene representation and novel view synthesis have witnessed the rise of Neural Radiance Fields (NeRFs). Nevertheless, it is not trivial to exploit NeRF for the photorealistic 3D scene stylization task, which aims to generate visually consistent and photorealistic stylized scenes from novel views. Simply coupling NeRF with photorealistic style transfer (PST) will result in cross-view inconsistency and degradation of stylized view syntheses. Through a thorough analysis, we demonstrate that this non-trivial task can be simplified in a new light: When transforming the appearance representation of a pre-trained NeRF with Lipschitz mapping, the consistency and photorealism across source views will be seamlessly encoded into the syntheses. That motivates us to build a concise and flexible learning framework namely LipRF, which upgrades arbitrary 2D PST methods with Lipschitz mapping tailored for the 3D scene. Technically, LipRF first pre-trains a radiance field to reconstruct the 3D scene, and then emulates the style on each view by 2D PST as the prior to learn a Lipschitz network to stylize the pre-trained appearance. In view of that Lipschitz condition highly impacts the expressivity of the neural network, we devise an adaptive regularization to balance the reconstruction and stylization. A gradual gradient aggregation strategy is further introduced to optimize LipRF in a cost-efficient manner. We conduct extensive experiments to show the high quality and robust performance of LipRF on both photorealistic 3D stylization and object appearance editing.
****
BEV-LaneDet: An Efficient 3D Lane Detection Based on Virtual Camera via Key-Points
Ruihao Wang, Jian Qin, Kaiying Li, Yaochen Li, Dong Cao, Jintao Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1002-1011
3D lane detection which plays a crucial role in vehicle routing, has recently been a rapidly developing topic in autonomous driving. Previous works struggle with practicality due to their complicated spatial transformations and inflexible representations of 3D lanes. Faced with the issues, our work proposes an efficien

t and robust monocular 3D lane detection called BEV-LaneDet with three main contributions. First, we introduce the Virtual Camera that unifies the in/extrinsic parameters of cameras mounted on different vehicles to guarantee the consistency of the spatial relationship among cameras. It can effectively promote the learning procedure due to the unified visual space. We secondly propose a simple but efficient 3D lane representation called Key-Points Representation. This module is more suitable to represent the complicated and diverse 3D lane structures. At last, we present a light-weight and chip-friendly spatial transformation module named Spatial Transformation Pyramid to transform multiscale front-view features into BEV features. Experimental results demonstrate that our work outperforms the state-of-the-art approaches in terms of F-Score, being 10.6% higher on the OpenLane dataset and 4.0% higher on the Apollo 3D synthetic dataset, with a speed of 185 FPS. Code is released at https://github.com/gigo-team/bev_lane_det.
****
Self-Supervised 3D Scene Flow Estimation Guided by Superpoints
Yaqi Shen, Le Hui, Jin Xie, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5271-5280
3D scene flow estimation aims to estimate point-wise motions between two consecutive frames of point clouds. Superpoints, i.e., points with similar geometric features, are usually employed to capture similar motions of local regions in 3D scenes for scene flow estimation. However, in existing methods, superpoints are generated with the offline clustering methods, which cannot characterize local regions with similar motions for complex 3D scenes well, leading to inaccurate scene flow estimation. To this end, we propose an iterative end-to-end superpoint based scene flow estimation framework, where the superpoints can be dynamically updated to guide the point-level flow prediction. Specifically, our framework consists of a flow guided superpoint generation module and a superpoint guided flow refinement module. In our superpoint generation module, we utilize the bidirectional flow information at the previous iteration to obtain the matching points of points and superpoint centers for soft point-to-superpoint association construction, in which the superpoints are generated for pairwise point clouds. With the generated superpoints, we first reconstruct the flow for each point by adaptively aggregating the superpoint-level flow, and then encode the consistency between the reconstructed flow of pairwise point clouds. Finally, we feed the consistency encoding along with the reconstructed flow into GRU to refine point-level flow. Extensive experiments on several different datasets show that our method can achieve promising performance.
****
DiffCollage: Parallel Generation of Large Content With Diffusion Models
Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, Ming-Yu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10188-10198
We present DiffCollage, a compositional diffusion model that can generate large content by leveraging diffusion models trained on generating pieces of the large content. Our approach is based on a factor graph representation where each factor node represents a portion of the content and a variable node represents their overlap. This representation allows us to aggregate intermediate outputs from diffusion models defined on individual nodes to generate content of arbitrary size and shape in parallel without resorting to an autoregressive generation procedure. We apply DiffCollage to various tasks, including infinite image generation, panorama image generation, and long-duration text-guided motion generation. Extensive experimental results with a comparison to strong autoregressive baselines verify the effectiveness of our approach.
****
Efficient Second-Order Plane Adjustment
Lipu Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13113-13121
Planes are generally used in 3D reconstruction for depth sensors, such as RGB-D cameras and LiDARs. This paper focuses on the problem of estimating the optimal planes and sensor poses to minimize the point-to-plane distance. The resulting l

east-squares problem is referred to as plane adjustment (PA) in the literature, which is the counterpart of bundle adjustment (BA) in visual reconstruction. Iterative methods are adopted to solve these least-squares problems. Typically, Newton's method is rarely used for a large-scale least-squares problem, due to the high computational complexity of the Hessian matrix. Instead, methods using an approximation of the Hessian matrix, such as the Levenberg-Marquardt (LM) method, are generally adopted. This paper adopts the Newton's method to efficiently solve the PA problem. Specifically, given poses, the optimal plane have a close-form solution. Thus we can eliminate planes from the cost function, which significantly reduces the number of variables. Furthermore, as the optimal planes are functions of poses, this method actually ensures that the optimal planes for the current estimated poses can be obtained at each iteration, which benefits the convergence. The difficulty lies in how to efficiently compute the Hessian matrix and the gradient of the resulting cost. This paper provides an efficient solution. Empirical evaluation shows that our algorithm outperforms the state-of-the-art algorithms.
****
Guided Depth Super-Resolution by Deep Anisotropic Diffusion
Nando Metzger, Rodrigo Caye Daudt, Konrad Schindler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18237-18246
Performing super-resolution of a depth image using the guidance from an RGB image is a problem that concerns several fields, such as robotics, medical imaging, and remote sensing. While deep learning methods have achieved good results in this problem, recent work highlighted the value of combining modern methods with more formal frameworks. In this work we propose a novel approach which combines guided anisotropic diffusion with a deep convolutional network and advances the state of the art for guided depth super-resolution. The edge transferring/enhancing properties of the diffusion are boosted by the contextual reasoning capabilities of modern networks, and a strict adjustment step guarantees perfect adherence to the source image. We achieve unprecedented results in three commonly used benchmarks for guided depth super resolution. The performance gain compared to other methods is the largest at larger scales, such as x32 scaling. Code for the proposed method will be made available to promote reproducibility of our results.
****
Fresnel Microfacet BRDF: Unification of Polari-Radiometric Surface-Body Reflection
Tomoki Ichikawa, Yoshiki Fukao, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16489-16497
Computer vision applications have heavily relied on the linear combination of Lambertian diffuse and microfacet specular reflection models for representing reflected radiance, which turns out to be physically incompatible and limited in applicability. In this paper, we derive a novel analytical reflectance model, which we refer to as Fresnel Microfacet BRDF model, that is physically accurate and generalizes to various real-world surfaces. Our key idea is to model the Fresnel reflection and transmission of the surface microgeometry with a collection of oriented mirror facets, both for body and surface reflections. We carefully derive the Fresnel reflection and transmission for each microfacet as well as the light transport between them in the subsurface. This physically-grounded modeling also allows us to express the polarimetric behavior of reflected light in addition to its radiometric behavior. That is, FMBRDF unifies not only body and surface reflections but also light reflection in radiometry and polarization and represents them in a single model. Experimental results demonstrate its effectiveness in accuracy, expressive power, image-based estimation, and geometry recovery.
****
A Unified Pyramid Recurrent Network for Video Frame Interpolation
Xin Jin, Longhai Wu, Jie Chen, Youxin Chen, Jayoon Koo, Cheul-hee Hahm; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1578-1587

Flow-guided synthesis provides a common framework for frame interpolation, where optical flow is estimated to guide the synthesis of intermediate frames between consecutive inputs. In this paper, we present UPR-Net, a novel Unified Pyramid Recurrent Network for frame interpolation. Cast in a flexible pyramid framework, UPR-Net exploits lightweight recurrent modules for both bi-directional flow estimation and intermediate frame synthesis. At each pyramid level, it leverages estimated bi-directional flow to generate forward-warped representations for frame synthesis; across pyramid levels, it enables iterative refinement for both optical flow and intermediate frame. In particular, we show that our iterative synthesis strategy can significantly improve the robustness of frame interpolation on large motion cases. Despite being extremely lightweight (1.7M parameters), our base version of UPR-Net achieves excellent performance on a large range of benchmarks. Code and trained models of our UPR-Net series are available at: https://github.com/srcn-ivl/UPR-Net.
****
Mofusion: A Framework for Denoising-Diffusion-Based Motion Synthesis
Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9760-9770
Conventional methods for human motion synthesis have either been deterministic or have had to struggle with the trade-off between motion diversity vs motion quality. In response to these limitations, we introduce MoFusion, i.e., a new denoising-diffusion-based framework for high-quality conditional human motion synthesis that can synthesise long, temporally plausible, and semantically accurate motions based on a range of conditioning contexts (such as music and text). We also present ways to introduce well-known kinematic losses for motion plausibility within the motion-diffusion framework through our scheduled weighting strategy. The learned latent space can be used for several interactive motion-editing applications like in-betweening, seed-conditioning, and text-based editing, thus, providing crucial abilities for virtual-character animation and robotics. Through comprehensive quantitative evaluations and a perceptual user study, we demonstrate the effectiveness of MoFusion compared to the state-of-the-art on established benchmarks in the literature. We urge the reader to watch our supplementary video. The source code will be released.
****
PoseFormerV2: Exploring Frequency Domain for Efficient and Robust 3D Human Pose Estimation
Qitao Zhao, Ce Zheng, Mengyuan Liu, Pichao Wang, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8877-8886
Recently, transformer-based methods have gained significant success in sequential 2D-to-3D lifting human pose estimation. As a pioneering work, PoseFormer captures spatial relations of human joints in each video frame and human dynamics across frames with cascaded transformer layers and has achieved impressive performance. However, in real scenarios, the performance of PoseFormer and its follow-ups is limited by two factors: (a) The length of the input joint sequence; (b) The quality of 2D joint detection. Existing methods typically apply self-attention to all frames of the input sequence, causing a huge computational burden when the frame number is increased to obtain advanced estimation accuracy, and they are not robust to noise naturally brought by the limited capability of 2D joint detectors. In this paper, we propose PoseFormerV2, which exploits a compact representation of lengthy skeleton sequences in the frequency domain to efficiently scale up the receptive field and boost robustness to noisy 2D joint detection. With minimum modifications to PoseFormer, the proposed method effectively fuses features both in the time domain and frequency domain, enjoying a better speed-accuracy trade-off than its precursor. Extensive experiments on two benchmark datasets (i.e., Human3.6M and MPI-INF-3DHP) demonstrate that the proposed approach significantly outperforms the original PoseFormer and other transformer-based variants. Code is released at https://github.com/QitaoZhao/PoseFormerV2.
****

Mask3D: Pre-Training 2D Vision Transformers by Learning Masked 3D Priors
Ji Hou, Xiaoliang Dai, Zijian He, Angela Dai, Matthias Nießner; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 13510-13519
Current popular backbones in computer vision, such as Vision Transformers (ViT)
and ResNets are trained to perceive the world from 2D images. However, to more e
ffectively understand 3D structural priors in 2D backbones, we propose Mask3D to
 leverage existing large-scale RGB-D data in a self-supervised pre-training to e
mbed these 3D priors into 2D learned feature representations. In contrast to tra
ditional 3D contrastive learning paradigms requiring 3D reconstructions or multi
-view correspondences, our approach is simple: we formulate a pre-text reconstru
ction task by masking RGB and depth patches in individual RGB-D frames. We demon
strate the Mask3D is particularly effective in embedding 3D priors into the powe
rful 2D ViT backbone, enabling improved representation learn- ing for various sc
ene understanding tasks, such as semantic segmentation, instance segmentation an
d object detection. Experiments show that Mask3D notably outperforms exist- ing
self-supervised 3D pre-training approaches on ScanNet, NYUv2, and Cityscapes ima
ge understanding tasks, with an improvement of +6.5% mIoU against the state-of-t
he-art Pri3D on ScanNet image semantic segmentation.
****
Physically Adversarial Infrared Patches With Learnable Shapes and Locations
Xingxing Wei, Jie Yu, Yao Huang; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2023, pp. 12334-12342
Owing to the extensive application of infrared object detectors in the safety-cr
itical tasks, it is necessary to evaluate their robustness against adversarial e
xamples in the real world. However, current few physical infrared attacks are co
mplicated to implement in practical application because of their complex transfo
rmation from digital world to physical world. To address this issue, in this pap
er, we propose a physically feasible infrared attack method called "adversarial
infrared patches". Considering the imaging mechanism of infrared cameras by capt
uring objects' thermal radiation, adversarial infrared patches conduct attacks b
y attaching a patch of thermal insulation materials on the target object to mani
pulate its thermal distribution. To enhance adversarial attacks, we present a no
vel aggregation regularization to guide the simultaneous learning for the patch'
 shape and location on the target object. Thus, a simple gradient-based optimiza
tion can be adapted to solve for them. We verify adversarial infrared patches in
 different object detection tasks with various object detectors. Experimental re
sults show that our method achieves more than 90% Attack Success Rate (ASR) vers
us the pedestrian detector and vehicle detector in the physical environment, whe
re the objects are captured in different angles, distances, postures, and scenes
. More importantly, adversarial infrared patch is easy to implement, and it only
 needs 0.5 hour to be constructed in the physical world, which verifies its effe
ctiveness and efficiency.
****
DiffusioNeRF: Regularizing Neural Radiance Fields With Denoising Diffusion Model
s
Jamie Wynn, Daniyar Turmukhambetov; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2023, pp. 4180-4189
Under good conditions, Neural Radiance Fields (NeRFs) have shown impressive resu
lts on novel view synthesis tasks. NeRFs learn a scene's color and density field
s by minimizing the photometric discrepancy between training views and different
iable renderings of the scene. Once trained from a sufficient set of views, NeRF
s can generate novel views from arbitrary camera positions. However, the scene g
eometry and color fields are severely under-constrained, which can lead to artif
acts, especially when trained with few input views. To alleviate this problem we
 learn a prior over scene geometry and color, using a denoising diffusion model
(DDM). Our DDM is trained on RGBD patches of the synthetic Hypersim dataset and
can be used to predict the gradient of the logarithm of a joint probability dist
ribution of color and depth patches. We show that, these gradients of logarithms
 of RGBD patch priors serve to regularize geometry and color of a scene. During

NeRF training, random RGBD patches are rendered and the estimated gradient of th
e log-likelihood is backpropagated to the color and density fields. Evaluations
on LLFF, the most relevant dataset, show that our learned prior achieves improve
d quality in the reconstructed geometry and improved generalization to novel vie
ws. Evaluations on DTU show improved reconstruction quality among NeRF methods.
****
Exemplar-FreeSOLO: Enhancing Unsupervised Instance Segmentation With Exemplars
Taoseef Ishtiak, Qing En, Yuhong Guo; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15424-15433
Instance segmentation seeks to identify and segment each object from images, whi
ch often relies on a large number of dense annotations for model training. To al
leviate this burden, unsupervised instance segmentation methods have been develo
ped to train class-agnostic instance segmentation models without any annotation.
 In this paper, we propose a novel unsupervised instance segmentation approach,
Exemplar-FreeSOLO, to enhance unsupervised instance segmentation by exploiting a
 limited number of unannotated and unsegmented exemplars. The proposed framework
 offers a new perspective on directly perceiving top-down information without an
notations. Specifically, Exemplar-FreeSOLO introduces a novel exemplarknowledge
abstraction module to acquire beneficial top-down guidance knowledge for instanc
es using unsupervised exemplar object extraction. Moreover, a new exemplar embed
ding contrastive module is designed to enhance the discriminative capability of
the segmentation model by exploiting the contrastive exemplar-based guidance kno
wledge in the embedding space. To evaluate the proposed ExemplarFreeSOLO, we con
duct comprehensive experiments and perform in-depth analyses on three image inst
ance segmentation datasets. The experimental results demonstrate that the propos
ed approach is effective and outperforms the state-of-the-art methods.
****
Multimodal Prompting With Missing Modalities for Visual Recognition
Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, Chen-Yu Lee; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14943
-14952
In this paper, we tackle two challenges in multimodal learning for visual recogn
ition: 1) when missing-modality occurs either during training or testing in real
-world situations; and 2) when the computation resources are not available to fi
netune on heavy transformer models. To this end, we propose to utilize prompt le
arning and mitigate the above two challenges together. Specifically, our modalit
y-missing-aware prompts can be plugged into multimodal transformers to handle ge
neral missing-modality cases, while only requiring less than 1% learnable parame
ters compared to training the entire model. We further explore the effect of dif
ferent prompt configurations and analyze the robustness to missing modality. Ext
ensive experiments are conducted to show the effectiveness of our prompt learnin
g framework that improves the performance under various missing-modality cases,
while alleviating the requirement of heavy model re-training. Code is available.
****
Edge-Aware Regional Message Passing Controller for Image Forgery Localization
Dong Li, Jiaying Zhu, Menglu Wang, Jiawei Liu, Xueyang Fu, Zheng-Jun Zha; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 8222-8232
Digital image authenticity has promoted research on image forgery localization.
Although deep learning-based methods achieve remarkable progress, most of them u
sually suffer from severe feature coupling between the forged and authentic regi
ons. In this work, we propose a two-step Edge-aware Regional Message Passing Con
trolling strategy to address the above issue. Specifically, the first step is to
 account for fully exploiting the edge information. It consists of two core desi
gns: context-enhanced graph construction and threshold-adaptive differentiable b
inarization edge algorithm. The former assembles the global semantic information
 to distinguish the features between the forged and authentic regions, while the
 latter stands on the output of the former to provide the learnable edges. In th
e second step, guided by the learnable edges, a region message passing controlle
r is devised to weaken the message passing between the forged and authentic regi

ons. In this way, our ERMPC is capable of explicitly modeling the inconsistency between the forged and authentic regions and enabling it to perform well on refined forged images. Extensive experiments on several challenging benchmarks show that our method is superior to state-of-the-art image forgery localization methods qualitatively and quantitatively.
****
Neural Koopman Pooling: Control-Inspired Temporal Dynamics Encoding for Skeleton-Based Action Recognition
Xinghan Wang, Xin Xu, Yadong Mu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10597-10607
Skeleton-based human action recognition is becoming increasingly important in a variety of fields. Most existing works train a CNN or GCN based backbone to extract spatial-temporal features, and use temporal average/max pooling to aggregate the information. However, these pooling methods fail to capture high-order dynamics information. To address the problem, we propose a plug-and-play module called Koopman pooling, which is a parameterized high-order pooling technique based on Koopman theory. The Koopman operator linearizes a non-linear dynamics system, thus providing a way to represent the complex system through the dynamics matrix, which can be used for classification. We also propose an eigenvalue normalization method to encourage the learned dynamics to be non-decaying and stable. Besides, we also show that our Koopman pooling framework can be easily extended to one-shot action recognition when combined with Dynamic Mode Decomposition. The proposed method is evaluated on three benchmark datasets, namely NTU RGB+D 60, 120 and NW-UCLA. Our experiments clearly demonstrate that Koopman pooling significantly improves the performance under both full-dataset and one-shot settings.
****
Simulated Annealing in Early Layers Leads to Better Generalization
Amir M. Sarfi, Zahra Karimpour, Muawiz Chaudhary, Nasir M. Khalid, Mirco Ravanelli, Sudhir Mudur, Eugene Belilovsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20205-20214
Recently, a number of iterative learning methods have been introduced to improve generalization. These typically rely on training for longer periods of time in exchange for improved generalization. LLF (later-layer-forgetting) is a state-of-the-art method in this category. It strengthens learning in early layers by periodically re-initializing the last few layers of the network. Our principal innovation in this work is to use Simulated annealing in EArly Layers (SEAL) of the network in place of re-initialization of later layers. Essentially, later layers go through the normal gradient descent process, while the early layers go through short stints of gradient ascent followed by gradient descent. Extensive experiments on the popular Tiny-ImageNet dataset benchmark and a series of transfer learning and few-shot learning tasks show that we outperform LLF by a significant margin. We further show that, compared to normal training, LLF features, although improving on the target task, degrade the transfer learning performance across all datasets we explored. In comparison, our method outperforms LLF across the same target datasets by a large margin. We also show that the prediction depth of our method is significantly lower than that of LLF and normal training, indicating on average better prediction performance.
****
Spatiotemporal Self-Supervised Learning for Point Clouds in the Wild
Yanhao Wu, Tong Zhang, Wei Ke, Sabine Süsstrunk, Mathieu Salzmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5251-5260
Self-supervised learning (SSL) has the potential to benefit many applications, particularly those where manually annotating data is cumbersome. One such situation is the semantic segmentation of point clouds. In this context, existing methods employ contrastive learning strategies and define positive pairs by performing various augmentation of point clusters in a single frame. As such, these methods do not exploit the temporal nature of LiDAR data. In this paper, we introduce an SSL strategy that leverages positive pairs in both the spatial and temporal domains. To this end, we design (i) a point-to-cluster learning strategy that ag

gregates spatial information to distinguish objects; and (ii) a cluster-to-cluster learning strategy based on unsupervised object tracking that exploits temporal correspondences. We demonstrate the benefits of our approach via extensive experiments performed by self-supervised training on two large-scale LiDAR datasets and transferring the resulting models to other point cloud segmentation benchmarks. Our results evidence that our method outperforms the state-of-the-art point cloud SSL methods.
****
Semi-Supervised Learning Made Simple With Self-Supervised Clustering
Enrico Fini, Pietro Astolfi, Karteek Alahari, Xavier Alameda-Pineda, Julien Mairal, Moin Nabi, Elisa Ricci; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3187-3197
Self-supervised learning models have been shown to learn rich visual representations without requiring human annotations. However, in many real-world scenarios, labels are partially available, motivating a recent line of work on semi-supervised methods inspired by self-supervised principles. In this paper, we propose a conceptually simple yet empirically powerful approach to turn clustering-based self-supervised methods such as SwAV or DINO into semi-supervised learners. More precisely, we introduce a multi-task framework merging a supervised objective using ground-truth labels and a self-supervised objective relying on clustering assignments with a single cross-entropy loss. This approach may be interpreted as imposing the cluster centroids to be class prototypes. Despite its simplicity, we provide empirical evidence that our approach is highly effective and achieves state-of-the-art performance on CIFAR100 and ImageNet.
****
Blind Image Quality Assessment via Vision-Language Correspondence: A Multitask Learning Perspective
Weixia Zhang, Guangtao Zhai, Ying Wei, Xiaokang Yang, Kede Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14071-14081
We aim at advancing blind image quality assessment (BIQA), which predicts the human perception of image quality without any reference information. We develop a general and automated multitask learning scheme for BIQA to exploit auxiliary knowledge from other tasks, in a way that the model parameter sharing and the loss weighting are determined automatically. Specifically, we first describe all candidate label combinations (from multiple tasks) using a textual template, and compute the joint probability from the cosine similarities of the visual-textual embeddings. Predictions of each task can be inferred from the joint distribution, and optimized by carefully designed loss functions. Through comprehensive experiments on learning three tasks - BIQA, scene classification, and distortion type identification, we verify that the proposed BIQA method 1) benefits from the scene classification and distortion type identification tasks and outperforms the state-of-the-art on multiple IQA datasets, 2) is more robust in the group maximum differentiation competition, and 3) realigns the quality annotations from different IQA datasets more effectively. The source code is available at https://github.com/zwx8981/LIQE.
****
Exploring Data Geometry for Continual Learning
Zhi Gao, Chen Xu, Feng Li, Yunde Jia, Mehrtash Harandi, Yuwei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24325-24334
Continual learning aims to efficiently learn from a non-stationary stream of data while avoiding forgetting the knowledge of old data. In many practical applications, data complies with non-Euclidean geometry. As such, the commonly used Euclidean space cannot gracefully capture non-Euclidean geometric structures of data, leading to inferior results. In this paper, we study continual learning from a novel perspective by exploring data geometry for the non-stationary stream of data. Our method dynamically expands the geometry of the underlying space to match growing geometric structures induced by new data, and prevents forgetting by keeping geometric structures of old data into account. In doing so, we make use

of the mixed-curvature space and propose an incremental search scheme, through which the growing geometric structures are encoded. Then, we introduce an angular-regularization loss and a neighbor-robustness loss to train the model, capable of penalizing the change of global geometric structures and local geometric structures. Experiments show that our method achieves better performance than baseline methods designed in Euclidean space.
****
Frequency-Modulated Point Cloud Rendering With Easy Editing
Yi Zhang, Xiaoyang Huang, Bingbing Ni, Teng Li, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 119-129
We develop an effective point cloud rendering pipeline for novel view synthesis, which enables high fidelity local detail reconstruction, real-time rendering and user-friendly editing. In the heart of our pipeline is an adaptive frequency modulation module called Adaptive Frequency Net (AFNet), which utilizes a hypernetwork to learn the local texture frequency encoding that is consecutively injected into adaptive frequency activation layers to modulate the implicit radiance signal. This mechanism improves the frequency expressive ability of the network with richer frequency basis support, only at a small computational budget. To further boost performance, a preprocessing module is also proposed for point cloud geometry optimization via point opacity estimation. In contrast to implicit rendering, our pipeline supports high-fidelity interactive editing based on point cloud manipulation. Extensive experimental results on NeRF-Synthetic, ScanNet, DTU and Tanks and Temples datasets demonstrate the superior performances achieved by our method in terms of PSNR, SSIM and LPIPS, in comparison to the state-of-the-art.
****
Integral Neural Networks
Kirill Solodskikh, Azim Kurbanov, Ruslan Aydarkhanov, Irina Zhelavskaya, Yury Parfenov, Dehua Song, Stamatios Lefkimmiatis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16113-16122
We introduce a new family of deep neural networks. Instead of the conventional representation of network layers as N-dimensional weight tensors, we use continuous layer representation along the filter and channel dimensions. We call such networks Integral Neural Networks (INNs). In particular, the weights of INNs are represented as continuous functions defined on N-dimensional hypercubes, and the discrete transformations of inputs to the layers are replaced by continuous integration operations, accordingly. During the inference stage, our continuous layers can be converted into the traditional tensor representation via numerical integral quadratures. Such kind of representation allows the discretization of a network to an arbitrary size with various discretization intervals for the integral kernels. This approach can be applied to prune the model directly on the edge device while featuring only a small performance loss at high rates of structural pruning without any fine-tuning. To evaluate the practical benefits of our proposed approach, we have conducted experiments using various neural network architectures for multiple tasks. Our reported results show that the proposed INNs achieve the same performance with their conventional discrete counterparts, while being able to preserve approximately the same performance (2 % accuracy loss for ResNet18 on Imagenet) at a high rate (up to 30%) of structural pruning without fine-tuning, compared to 65 % accuracy loss of the conventional pruning methods under the same conditions.
****
Learning Neural Parametric Head Models
Simon Giebenhain, Tobias Kirschstein, Markos Georgopoulos, Martin Rünz, Lourdes Agapito, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21003-21012
We propose a novel 3D morphable model for complete human heads based on hybrid neural fields. At the core of our model lies a neural parametric representation that disentangles identity and expressions in disjoint latent spaces. To this end, we capture a person's identity in a canonical space as a signed distance field

 (SDF), and model facial expressions with a neural deformation field. In additio
n, our representation achieves high-fidelity local detail by introducing an ense
mble of local fields centered around facial anchor points. To facilitate general
ization, we train our model on a newly-captured dataset of over 3700 head scans
from 203 different identities using a custom high-end 3D scanning setup. Our dat
aset significantly exceeds comparable existing datasets, both with respect to qu
ality and completeness of geometry, averaging around 3.5M mesh faces per scan. F
inally, we demonstrate that our approach outperforms state-of-the-art methods in
 terms of fitting error and reconstruction quality.
****
Removing Objects From Neural Radiance Fields
Silvan Weder, Guillermo Garcia-Hernando, Áron Monszpart, Marc Pollefeys, Gabriel
 J. Brostow, Michael Firman, Sara Vicente; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16528-16538
Neural Radiance Fields (NeRFs) are emerging as a ubiquitous scene representation
 that allows for novel view synthesis. Increasingly, NeRFs will be shareable wit
h other people. Before sharing a NeRF, though, it might be desirable to remove p
ersonal information or unsightly objects. Such removal is not easily achieved wi
th the current NeRF editing frameworks. We propose a framework to remove objects
 from a NeRF representation created from an RGB-D sequence. Our NeRF inpainting
method leverages recent work in 2D image inpainting and is guided by a user-prov
ided mask. Our algorithm is underpinned by a confidence based view selection pro
cedure. It chooses which of the individual 2D inpainted images to use in the cre
ation of the NeRF, so that the resulting inpainted NeRF is 3D consistent. We sho
w that our method for NeRF editing is effective for synthesizing plausible inpai
ntings in a multi-view coherent manner, outperforming competing methods. We vali
date our approach by proposing a new and still-challenging dataset for the task
of NeRF inpainting.
****
Few-Shot Referring Relationships in Videos
Yogesh Kumar, Anand Mishra; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2023, pp. 2289-2298
Interpreting visual relationships is a core aspect of comprehensive video unders
tanding. Given a query visual relationship as <subject, predicate, object> and a
 test video, our objective is to localize the subject and object that are connec
ted via the predicate. Given modern visio-lingual understanding capabilities, so
lving this problem is achievable, provided that there are large-scale annotated
training examples available. However, annotating for every combination of subjec
t, object, and predicate is cumbersome, expensive, and possibly infeasible. Ther
efore, there is a need for models that can learn to spatially and temporally loc
alize subjects and objects that are connected via an unseen predicate using only
 a few support set videos sharing the common predicate. We address this challeng
ing problem, referred to as few-shot referring relationships in videos for the f
irst time. To this end, we pose the problem as a minimization of an objective fu
nction defined over a T-partite random field. Here, the vertices of the random f
ield correspond to candidate bounding boxes for the subject and object, and T re
presents the number of frames in the test video. This objective function is comp
osed of frame level and visual relationship similarity potentials. To learn thes
e potentials, we use a relation network that takes query-conditioned translation
al relationship embedding as inputs and is meta-trained using support set videos
 in an episodic manner. Further, the objective function is minimized using a bel
ief propagation-based message passing on the random field to obtain the spatiote
mporal localization or subject and object trajectories. We perform extensive exp
eriments using two public benchmarks, namely ImageNet-VidVRD and VidOR, and comp
are the proposed approach with competitive baselines to assess its efficacy.
****
Structural Multiplane Image: Bridging Neural View Synthesis and 3D Reconstructio
n
Mingfang Zhang, Jinglu Wang, Xiao Li, Yifei Huang, Yoichi Sato, Yan Lu; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)

, 2023, pp. 16707-16716
The Multiplane Image (MPI), containing a set of fronto-parallel RGBA layers, is an effective and efficient representation for view synthesis from sparse inputs. Yet, its fixed structure limits the performance, especially for surfaces imaged at oblique angles. We introduce the Structural MPI (S-MPI), where the plane structure approximates 3D scenes concisely. Conveying RGBA contexts with geometrically-faithful structures, the S-MPI directly bridges view synthesis and 3D reconstruction. It can not only overcome the critical limitations of MPI, i.e., discretization artifacts from sloped surfaces and abuse of redundant layers, and can also acquire planar 3D reconstruction. Despite the intuition and demand of applying S-MPI, great challenges are introduced, e.g., high-fidelity approximation for both RGBA layers and plane poses, multi-view consistency, non-planar regions modeling, and efficient rendering with intersected planes. Accordingly, we propose a transformer-based network based on a segmentation model. It predicts compact and expressive S-MPI layers with their corresponding masks, poses, and RGBA contexts. Non-planar regions are inclusively handled as a special case in our unified framework. Multi-view consistency is ensured by sharing global proxy embeddings, which encode plane-level features covering the complete 3D scenes with aligned coordinates. Intensive experiments show that our method outperforms both previous state-of-the-art MPI-based view synthesis methods and planar reconstruction methods.
****
Harmonious Teacher for Cross-Domain Object Detection
Jinhong Deng, Dongli Xu, Wen Li, Lixin Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23829-23838
Self-training approaches recently achieved promising results in cross-domain object detection, where people iteratively generate pseudo labels for unlabeled target domain samples with a model, and select high-confidence samples to refine the model. In this work, we reveal that the consistency of classification and localization predictions are crucial to measure the quality of pseudo labels, and propose a new Harmonious Teacher approach to improve the self-training for cross-domain object detection. In particular, we first propose to enhance the quality of pseudo labels by regularizing the consistency of the classification and localization scores when training the detection model. The consistency losses are defined for both labeled source samples and the unlabeled target samples. Then, we further remold the traditional sample selection method by a sample reweighing strategy based on the consistency of classification and localization scores to improve the ranking of predictions. This allows us to fully exploit all instance predictions from the target domain without abandoning valuable hard examples. Without bells and whistles, our method shows superior performance in various cross-domain scenarios compared with the state-of-the-art baselines, which validates the effectiveness of our Harmonious Teacher. Our codes will be available at https://github.com/kinredon/Harmonious-Teacher.
****
3D Human Pose Estimation via Intuitive Physics
Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Omid Taheri, Michael J. Black, Dimitrios Tzionas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4713-4725
Estimating 3D humans from images often produces implausible bodies that lean, float, or penetrate the floor. Such methods ignore the fact that bodies are typically supported by the scene. A physics engine can be used to enforce physical plausibility, but these are not differentiable, rely on unrealistic proxy bodies, and are difficult to integrate into existing optimization and learning frameworks. In contrast, we exploit novel intuitive-physics (IP) terms that can be inferred from a 3D SMPL body interacting with the scene. Inspired by biomechanics, we infer the pressure heatmap on the body, the Center of Pressure (CoP) from the heatmap, and the SMPL body's Center of Mass (CoM). With these, we develop IPMAN, to estimate a 3D body from a color image in a "stable" configuration by encouraging plausible floor contact and overlapping CoP and CoM. Our IP terms are intuitive, easy to implement, fast to compute, differentiable, and can be integrated int

o existing optimization and regression methods. We evaluate IPMAN on standard da
tasets and MoYo, a new dataset with synchronized multi-view images, ground-truth
 3D bodies with complex poses, body-floor contact, CoM and pressure. IPMAN produ
ces more plausible results than the state of the art, improving accuracy for sta
tic poses, while not hurting dynamic ones. Code and data are available for resea
rch at https://ipman.is.tue.mpg.de/.
****
SplineCam: Exact Visualization and Characterization of Deep Network Geometry and
 Decision Boundaries
Ahmed Imtiaz Humayun, Randall Balestriero, Guha Balakrishnan, Richard G. Baraniu
k; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2023, pp. 3789-3798
Current Deep Network (DN) visualization and interpretability methods rely heavil
y on data space visualizations such as scoring which dimensions of the data are
responsible for their associated prediction or generating new data features or s
amples that best match a given DN unit or representation. In this paper, we go o
ne step further by developing the first provably exact method for computing the
geometry of a DN's mapping -- including its decision boundary -- over a specifie
d region of the data space. By leveraging the theory of Continuous Piecewise Lin
ear (CPWL) spline DNs, SplineCam exactly computes a DN's geometry without resort
ing to approximations such as sampling or architecture simplification. SplineCam
 applies to any DN architecture based on CPWL activation nonlinearities, includi
ng (leaky) ReLU, absolute value, maxout, and max-pooling and can also be applied
 to regression DNs such as implicit neural representations. Beyond decision boun
dary visualization and characterization, SplineCam enables one to compare archit
ectures, measure generalizability, and sample from the decision boundary on or o
ff the data manifold. Project website: https://bit.ly/splinecam
****
Learning To Predict Scene-Level Implicit 3D From Posed RGBD Data
Nilesh Kulkarni, Linyi Jin, Justin Johnson, David F. Fouhey; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
 17256-17265
We introduce a method that can learn to predict scene-level implicit functions f
or 3D reconstruction from posed RGBD data. At test time, our system maps a previ
ously unseen RGB image to a 3D reconstruction of a scene via implicit functions.
 While implicit functions for 3D reconstruction have often been tied to meshes,
we show that we can train one using only a set of posed RGBD images. This settin
g may help 3D reconstruction unlock the sea of accelerometer+RGBD data that is c
oming with new phones. Our system, D2-DRDF, can match and sometimes outperform c
urrent methods that use mesh supervision and shows better robustness to sparse d
ata.
****
EXCALIBUR: Encouraging and Evaluating Embodied Exploration
Hao Zhu, Raghav Kapoor, So Yeon Min, Winson Han, Jiatai Li, Kaiwen Geng, Graham
Neubig, Yonatan Bisk, Aniruddha Kembhavi, Luca Weihs; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14931-
14942
Experience precedes understanding. Humans constantly explore and learn about the
ir environment out of curiosity, gather information, and update their models of
the world. On the other hand, machines are either trained to learn passively fro
m static and fixed datasets, or taught to complete specific goal-conditioned tas
ks. To encourage the development of exploratory interactive agents, we present t
he EXCALIBUR benchmark. EXCALIBUR allows agents to explore their environment for
 long durations and then query their understanding of the physical world via inq
uiries like: "is the small heavy red bowl made from glass?" or "is there a silve
r spoon heavier than the egg?". This design encourages agents to perform free-fo
rm home exploration without myopia induced by goal conditioning. Once the agents
 have answered a series of questions, they can renter the scene to refine their
knowledge, update their beliefs, and improve their performance on the questions.
 Our experiments demonstrate the challenges posed by this dataset for the presen

t-day state-of-the-art embodied systems and the headroom afforded to develop new innovative methods. Finally, we present a virtual reality interface that enables humans to seamlessly interact within the simulated world and use it to gather human performance measures. EXCALIBUR affords unique challenges in comparison to present-day benchmarks and represents the next frontier for embodied AI research.
****
Visual DNA: Representing and Comparing Images Using Distributions of Neuron Activations
Benjamin Ramtoula, Matthew Gadd, Paul Newman, Daniele De Martini; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11113-11123
Selecting appropriate datasets is critical in modern computer vision. However, no general-purpose tools exist to evaluate the extent to which two datasets differ. For this, we propose representing images -- and by extension datasets -- using Distributions of Neuron Activations (DNAs). DNAs fit distributions, such as histograms or Gaussians, to activations of neurons in a pre-trained feature extractor through which we pass the image(s) to represent. This extractor is frozen for all datasets, and we rely on its generally expressive power in feature space. By comparing two DNAs, we can evaluate the extent to which two datasets differ with granular control over the comparison attributes of interest, providing the ability to customise the way distances are measured to suit the requirements of the task at hand. Furthermore, DNAs are compact, representing datasets of any size with less than 15 megabytes. We demonstrate the value of DNAs by evaluating their applicability on several tasks, including conditional dataset comparison, synthetic image evaluation, and transfer learning, and across diverse datasets, ranging from synthetic cat images to celebrity faces and urban driving scenes.
****
Recognizability Embedding Enhancement for Very Low-Resolution Face Recognition and Quality Estimation
Jacky Chen Long Chai, Tiong-Sik Ng, Cheng-Yaw Low, Jaewoo Park, Andrew Beng Jin Teoh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9957-9967
Very low-resolution face recognition (VLRFR) poses unique challenges, such as tiny regions of interest and poor resolution due to extreme standoff distance or wide viewing angle of the acquisition device. In this paper, we study principled approaches to elevate the recognizability of a face in the embedding space instead of the visual quality. We first formulate a robust learning-based face recognizability measure, namely recognizability index (RI), based on two criteria: (i) proximity of each face embedding against the unrecognizable faces cluster center and (ii) closeness of each face embedding against its positive and negative class prototypes. We then devise an index diversion loss to push the hard-to-recognize face embedding with low RI away from unrecognizable faces cluster to boost the RI, which reflects better recognizability. Additionally, a perceptibility-aware attention mechanism is introduced to attend to the salient recognizable face regions, which offers better explanatory and discriminative content for embedding learning. Our proposed model is trained end-to-end and simultaneously serves recognizability-aware embedding learning and face quality estimation. To address VLRFR, extensive evaluations on three challenging low-resolution datasets and face quality assessment demonstrate the superiority of the proposed model over the state-of-the-art methods.
****
Physical-World Optical Adversarial Attacks on 3D Face Recognition
Yanjie Li, Yiquan Li, Xuelong Dai, Songtao Guo, Bin Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24699-24708
The success rate of current adversarial attacks remains low on real-world 3D face recognition tasks because the 3D-printing attacks need to meet the requirement that the generated points should be adjacent to the surface, which limits the adversarial example' searching space. Additionally, they have not considered unpr

edictable head movements or the non-homogeneous nature of skin reflectance in th e real world. To address the real-world challenges, we propose a novel structure d-light attack against structured-light-based 3D face recognition. We incorporat e the 3D reconstruction process and skin's reflectance in the optimization proce ss to get the end-to-end attack and present 3D transform invariant loss and sens itivity maps to improve robustness. Our attack enables adversarial points to be placed in any position and is resilient to random head movements while maintaini ng the perturbation unnoticeable. Experiments show that our new method can attac k point-cloud-based and depth-image-based 3D face recognition systems with a hig h success rate, using fewer perturbations than previous physical 3D adversarial attacks.
****

Accelerating Dataset Distillation via Model Augmentation
Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwe n Ding, Yao Li, Dongkuan Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11950-11959
Dataset Distillation (DD), a newly emerging field, aims at generating much small er but efficient synthetic training datasets from large ones. Existing DD method s based on gradient matching achieve leading performance; however, they are extr emely computationally intensive as they require continuously optimizing a datase t among thousands of randomly initialized models. In this paper, we assume that training the synthetic data with diverse models leads to better generalization p erformance. Thus we propose two model augmentation techniques, i.e. using early-stage models and parameter perturbation to learn an informative synthetic set wi th significantly reduced training cost. Extensive experiments demonstrate that o ur method achieves up to 20x speedup and comparable performance on par with stat e-of-the-art methods.
****

SE-ORNet: Self-Ensembling Orientation-Aware Network for Unsupervised Point Cloud Shape Correspondence
Jiacheng Deng, Chuxin Wang, Jiahao Lu, Jianfeng He, Tianzhu Zhang, Jiyang Yu, Zh e Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2023, pp. 5364-5373
Unsupervised point cloud shape correspondence aims to obtain dense point-to-poin t correspondences between point clouds without manually annotated pairs. However , humans and some animals have bilateral symmetry and various orientations, whic h leads to severe mispredictions of symmetrical parts. Besides, point cloud nois e disrupts consistent representations for point cloud and thus degrades the shap e correspondence accuracy. To address the above issues, we propose a Self-Ensemb ling ORientation-aware Network termed SE-ORNet. The key of our approach is to ex ploit an orientation estimation module with a domain adaptive discriminator to a lign the orientations of point cloud pairs, which significantly alleviates the m ispredictions of symmetrical parts. Additionally, we design a self-ensembling fr amework for unsupervised point cloud shape correspondence. In this framework, th e disturbances of point cloud noise are overcome by perturbing the inputs of the student and teacher networks with different data augmentations and constraining the consistency of predictions. Extensive experiments on both human and animal datasets show that our SE-ORNet can surpass state-of-the-art unsupervised point cloud shape correspondence methods.
****

Raw Image Reconstruction With Learned Compact Metadata
Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C. Kot, Bihan We n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2023, pp. 18206-18215
While raw images exhibit advantages over sRGB images (e.g. linearity and fine-gr ained quantization level), they are not widely used by common users due to the l arge storage requirements. Very recent works propose to compress raw images by d esigning the sampling masks in the raw image pixel space, leading to suboptimal image representations and redundant metadata. In this paper, we propose a novel framework to learn a compact representation in the latent space serving as the m

etadata in an end-to-end manner. Furthermore, we propose a novel sRGB-guided context model with the improved entropy estimation strategies, which leads to better reconstruction quality, smaller size of metadata, and faster speed. We illustrate how the proposed raw image compression scheme can adaptively allocate more bits to image regions that are important from a global perspective. The experimental results show that the proposed method can achieve superior raw image reconstruction results using a smaller size of the metadata on both uncompressed sRGB images and JPEG images.
****

Semi-Supervised Video Inpainting With Cycle Consistency Constraints

Zhiliang Wu, Hanyu Xuan, Changchang Sun, Weili Guan, Kang Zhang, Yan Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22586-22595

Deep learning-based video inpainting has yielded promising results and gained increasing attention from researchers. Generally, these methods usually assume that the corrupted region masks of each frame are known and easily obtained. However, the annotation of these masks are labor-intensive and expensive, which limits the practical application of current methods. Therefore, we expect to relax this assumption by defining a new semi-supervised inpainting setting, making the networks have the ability of completing the corrupted regions of the whole video using the annotated mask of only one frame. Specifically, in this work, we propose an end-to-end trainable framework consisting of completion network and mask prediction network, which are designed to generate corrupted contents of the current frame using the known mask and decide the regions to be filled of the next frame, respectively. Besides, we introduce a cycle consistency loss to regularize the training parameters of these two networks. In this way, the completion network and the mask prediction network can constrain each other, and hence the overall performance of the trained model can be maximized. Furthermore, due to the natural existence of prior knowledge (e.g., corrupted contents and clear borders), current video inpainting datasets are not suitable in the context of semi-supervised video inpainting. Thus, we create a new dataset by simulating the corrupted video of real-world scenarios. Extensive experimental results are reported to demonstrate the superiority of our model in the video inpainting task. Remarkably, although our model is trained in a semi-supervised manner, it can achieve comparable performance as fully-supervised methods.
****

Frame-Event Alignment and Fusion Network for High Frame Rate Tracking

Jiqing Zhang, Yuanchen Wang, Wenxi Liu, Meng Li, Jinpeng Bai, Baocai Yin, Xin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9781-9790

Most existing RGB-based trackers target low frame rate benchmarks of around 30 frames per second. This setting restricts the tracker's functionality in the real world, especially for fast motion. Event-based cameras as bioinspired sensors provide considerable potential for high frame rate tracking due to their high temporal resolution. However, event-based cameras cannot offer fine-grained texture information like conventional cameras. This unique complementarity motivates us to combine conventional frames and events for high frame rate object tracking under various challenging conditions. In this paper, we propose an end-to-end network consisting of multi-modality alignment and fusion modules to effectively combine meaningful information from both modalities at different measurement rates. The alignment module is responsible for cross-modality and cross-frame-rate alignment between frame and event modalities under the guidance of the moving cues furnished by events. While the fusion module is accountable for emphasizing valuable features and suppressing noise information by the mutual complement between the two modalities. Extensive experiments show that the proposed approach outperforms state-of-the-art trackers by a significant margin in high frame rate tracking. With the FE240hz dataset, our approach achieves high frame rate tracking up to 240Hz.
****

A Bag-of-Prototypes Representation for Dataset-Level Applications

Weijie Tu, Weijian Deng, Tom Gedeon, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2881-2892
This work investigates dataset vectorization for two dataset-level tasks: assessing training set suitability and test set difficulty. The former measures how suitable a training set is for a target domain, while the latter studies how challenging a test set is for a learned model. Central of the two tasks is measuring the underlying relationship between datasets. This needs a desirable dataset vectorization scheme, which should preserve as much discriminative dataset information as possible so that the distance between the resulting dataset vectors can reflect dataset-to-dataset similarity. To this end, we propose a bag-of-prototypes (BoP) dataset representation that extends the image level bag consisting of patch descriptors to dataset-level bag consisting of semantic prototypes. Specifically, we develop a codebook consisting of K prototypes clustered from a reference dataset. Given a dataset to be encoded, we quantize each of its image features to a certain prototype in the codebook and obtain a K-dimensional histogram feature. Without assuming access to dataset labels, the BoP representation provides rich characterization of dataset semantic distribution. Further, BoP representations cooperates well with Jensen-Shannon divergence for measuring dataset-to-dataset similarity. Albeit very simple, BoP consistently shows its advantage over existing representations on a series of benchmarks for two dataset-level tasks.
****
Level-S^{2}fM: Structure From Motion on Neural Level Set of Implicit Surfaces
Yuxi Xiao, Nan Xue, Tianfu Wu, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17205-17214
This paper presents a neural incremental Structure-from-Motion (SfM) approach, Level-S2fM, which estimates the camera poses and scene geometry from a set of uncalibrated images by learning coordinate MLPs for the implicit surfaces and the radiance fields from the established keypoint correspondences. Our novel formulation poses some new challenges due to inevitable two-view and few-view configurations in the incremental SfM pipeline, which complicates the optimization of coordinate MLPs for volumetric neural rendering with unknown camera poses. Nevertheless, we demonstrate that the strong inductive basis conveying in the 2D correspondences is promising to tackle those challenges by exploiting the relationship between the ray sampling schemes. Based on this, we revisit the pipeline of incremental SfM and renew the key components, including two-view geometry initialization, the camera poses registration, the 3D points triangulation, and Bundle Adjustment, with a fresh perspective based on neural implicit surfaces. By unifying the scene geometry in small MLP networks through coordinate MLPs, our Level-S2fM treats the zero-level set of the implicit surface as an informative top-down regularization to manage the reconstructed 3D points, reject the outliers in correspondences via querying SDF, and refine the estimated geometries by NBA (Neural BA). Not only does our Level-S2fM lead to promising results on camera pose estimation and scene geometry reconstruction, but it also shows a promising way for neural implicit rendering without knowing camera extrinsic beforehand.
****
Neuron Structure Modeling for Generalizable Remote Physiological Measurement
Hao Lu, Zitong Yu, Xuesong Niu, Ying-Cong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18589-18599
Remote photoplethysmography (rPPG) technology has drawn increasing attention in recent years. It can extract Blood Volume Pulse (BVP) from facial videos, making many applications like health monitoring and emotional analysis more accessible. However, as the BVP signal is easily affected by environmental changes, existing methods struggle to generalize well for unseen domains. In this paper, we systematically address the domain shift problem in the rPPG measurement task. We show that most domain generalization methods do not work well in this problem, as domain labels are ambiguous in complicated environmental changes. In light of this, we propose a domain-label-free approach called NEuron STructure modeling (NEST). NEST improves the generalization capacity by maximizing the coverage of feature space during training, which reduces the chance for under-optimized feature activation during inference. Besides, NEST can also enrich and enhance domain i

nvariant features across multi-domain. We create and benchmark a large-scale domain generalization protocol for the rPPG measurement task. Extensive experiments show that our approach outperforms the state-of-the-art methods on both cross-dataset and intra-dataset settings.
****
Shape-Aware Text-Driven Layered Video Editing
Yao-Chih Lee, Ji-Ze Genevieve Jang, Yi-Ting Chen, Elizabeth Qiu, Jia-Bin Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14317-14326
Temporal consistency is essential for video editing applications. Existing work on layered representation of videos allows propagating edits consistently to each frame. These methods, however, can only edit object appearance rather than object shape changes due to the limitation of using a fixed UV mapping field for texture atlas. We present a shape-aware, text-driven video editing method to tackle this challenge. To handle shape changes in video editing, we first propagate the deformation field between the input and edited keyframe to all frames. We then leverage a pre-trained text-conditioned diffusion model as guidance for refining shape distortion and completing unseen regions. The experimental results demonstrate that our method can achieve shape-aware consistent video editing and compare favorably with the state-of-the-art.
****
Out-of-Candidate Rectification for Weakly Supervised Semantic Segmentation
Zesen Cheng, Pengchong Qiao, Kehan Li, Siheng Li, Pengxu Wei, Xiangyang Ji, Li Yuan, Chang Liu, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23673-23684
Weakly supervised semantic segmentation is typically inspired by class activation maps, which serve as pseudo masks with class-discriminative regions highlighted. Although tremendous efforts have been made to recall precise and complete locations for each class, existing methods still commonly suffer from the unsolicited Out-of-Candidate (OC) error predictions that do not belong to the label candidates, which could be avoidable since the contradiction with image-level class tags is easy to be detected. In this paper, we develop a group ranking-based Out-of-Candidate Rectification (OCR) mechanism in a plug-and-play fashion. Firstly, we adaptively split the semantic categories into In-Candidate (IC) and OC groups for each OC pixel according to their prior annotation correlation and posterior prediction correlation. Then, we derive a differentiable rectification loss to force OC pixels to shift to the IC group. Incorporating OCR with seminal baselines (e.g., AffinityNet, SEAM, MCTformer), we can achieve remarkable performance gains on both Pascal VOC (+3.2%, +3.3%, +0.8% mIoU) and MS COCO (+1.0%, +1.3%, +0.5% mIoU) datasets with negligible extra training overhead, which justifies the effectiveness and generality of OCR.
****
Solving Relaxations of MAP-MRF Problems: Combinatorial In-Face Frank-Wolfe Directions
Vladimir Kolmogorov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11980-11989
We consider the problem of solving LP relaxations of MAP-MRF inference problems, and in particular the method proposed recently in (Swoboda, Kolmogorov 2019; Kolmogorov, Pock 2021). As a key computational subroutine, it uses a variant of the Frank-Wolfe (FW) method to minimize a smooth convex function over a combinatorial polytope. We propose an efficient implementation of this subproutine based on in-face Frank-Wolfe directions, introduced in (Freund et al. 2017) in a different context. More generally, we define an abstract data structure for a combinatorial subproblem that enables in-face FW directions, and describe its specialization for tree-structured MAP-MRF inference subproblems. Experimental results indicate that the resulting method is the current state-of-art LP solver for some classes of problems. Our code is available at pub.ist.ac.at/ vnk/papers/IN-FACE-FW.html.
****
MEGANE: Morphable Eyeglass and Avatar Network

Junxuan Li, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Hongdong Li, Jason Saragih; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12769-12779
Eyeglasses play an important role in the perception of identity. Authentic virtual representations of faces can benefit greatly from their inclusion. However, modeling the geometric and appearance interactions of glasses and the face of virtual representations of humans is challenging. Glasses and faces affect each other's geometry at their contact points, and also induce appearance changes due to light transport. Most existing approaches do not capture these physical interactions since they model eyeglasses and faces independently. Others attempt to resolve interactions as a 2D image synthesis problem and suffer from view and temporal inconsistencies. In this work, we propose a 3D compositional morphable model of eyeglasses that accurately incorporates high-fidelity geometric and photometric interaction effects. To support the large variation in eyeglass topology efficiently, we employ a hybrid representation that combines surface geometry and a volumetric representation. Unlike volumetric approaches, our model naturally retains correspondences across glasses, and hence explicit modification of geometry, such as lens insertion and frame deformation, is greatly simplified. In addition, our model is relightable under point lights and natural illumination, supporting high-fidelity rendering of various frame materials, including translucent plastic and metal within a single morphable model. Importantly, our approach models global light transport effects, such as casting shadows between faces and glasses. Our morphable model for eyeglasses can also be fit to novel glasses via inverse rendering. We compare our approach to state-of-the-art methods and demonstrate significant quality improvements.
****
Leverage Interactive Affinity for Affordance Learning
Hongchen Luo, Wei Zhai, Jing Zhang, Yang Cao, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6809-6819
Perceiving potential "action possibilities" (i.e., affordance) regions of images and learning interactive functionalities of objects from human demonstration is a challenging task due to the diversity of human-object interactions. Prevailing affordance learning algorithms often adopt the label assignment paradigm and presume that there is a unique relationship between functional region and affordance label, yielding poor performance when adapting to unseen environments with large appearance variations. In this paper, we propose to leverage interactive affinity for affordance learning, i.e., extracting interactive affinity from human-object interaction and transferring it to non-interactive objects. Interactive affinity, which represents the contacts between different parts of the human body and local regions of the target object, can provide inherent cues of interconnectivity between humans and objects, thereby reducing the ambiguity of the perceived action possibilities. Specifically, we propose a pose-aided interactive affinity learning framework that exploits human pose to guide the network to learn the interactive affinity from human-object interactions. Particularly, a keypoint heuristic perception (KHP) scheme is devised to exploit the keypoint association of human pose to alleviate the uncertainties due to interaction diversities and contact occlusions. Besides, a contact-driven affordance learning (CAL) dataset is constructed by collecting and labeling over 5,000 images. Experimental results demonstrate that our method outperforms the representative models regarding objective metrics and visual quality. Code and dataset: github.com/lhc1224/PIAL-Net.
****
Enhancing Multiple Reliability Measures via Nuisance-Extended Information Bottleneck
Jongheon Jeong, Sihyun Yu, Hankook Lee, Jinwoo Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16206-16218
In practical scenarios where training data is limited, many predictive signals in the data can be rather from some biases in data acquisition (i.e., less genera

lizable), so that one cannot prevent a model from co-adapting on such (so-called) "shortcut" signals: this makes the model fragile in various distribution shifts. To bypass such failure modes, we consider an adversarial threat model under a mutual information constraint to cover a wider class of perturbations in training. This motivates us to extend the standard information bottleneck to additionally model the nuisance information. We propose an autoencoder-based training to implement the objective, as well as practical encoder designs to facilitate the proposed hybrid discriminative-generative training concerning both convolutional- and Transformer-based architectures. Our experimental results show that the proposed scheme improves robustness of learned representations (remarkably without using any domain-specific knowledge), with respect to multiple challenging reliability measures. For example, our model could advance the state-of-the-art on a recent challenging OBJECTS benchmark in novelty detection by 78.4% -> 87.2% in AUROC, while simultaneously enjoying improved corruption, background and (certified) adversarial robustness. Code is available at https://github.com/jh-jeong/nuisance_ib.
****
Rethinking the Approximation Error in 3D Surface Fitting for Point Cloud Normal Estimation
Hang Du, Xuejun Yan, Jingjing Wang, Di Xie, Shiliang Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9486-9495
Most existing approaches for point cloud normal estimation aim to locally fit a geometric surface and calculate the normal from the fitted surface. Recently, learning-based methods have adopted a routine of predicting point-wise weights to solve the weighted least-squares surface fitting problem. Despite achieving remarkable progress, these methods overlook the approximation error of the fitting problem, resulting in a less accurate fitted surface. In this paper, we first carry out in-depth analysis of the approximation error in the surface fitting problem. Then, in order to bridge the gap between estimated and precise surface normals, we present two basic design principles: 1) applies the Z-direction Transform to rotate local patches for a better surface fitting with a lower approximation error; 2) models the error of the normal estimation as a learnable term. We implement these two principles using deep neural networks, and integrate them with the state-of-the-art (SOTA) normal estimation methods in a plug-and-play manner. Extensive experiments verify our approaches bring benefits to point cloud normal estimation and push the frontier of state-of-the-art performance on both synthetic and real-world datasets. The code is available at https://github.com/hikvision-research/3DVision.
****
Objaverse: A Universe of Annotated 3D Objects
Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, Ali Farhadi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13142-13153
Massive data corpora like WebText, Wikipedia, Conceptual Captions, WebImageText, and LAION have propelled recent dramatic progress in AI. Large neural models trained on such datasets produce impressive results and top many of today's benchmarks. A notable omission within this family of large-scale datasets is 3D data. Despite considerable interest and potential applications in 3D vision, datasets of high-fidelity 3D models continue to be mid-sized with limited diversity of object categories. Addressing this gap, we present Objaverse 1.0, a large dataset of objects with 800K+ (and growing) 3D models with descriptive captions, tags, and animations. Objaverse improves upon present day 3D repositories in terms of scale, number of categories, and in the visual diversity of instances within a category. We demonstrate the large potential of Objaverse via four diverse applications: training generative 3D models, improving tail category segmentation on the LVIS benchmark, training open-vocabulary object-navigation models for Embodied AI, and creating a new benchmark for robustness analysis of vision models. Objaverse can open new directions for research and enable new applications across th

e field of AI.
****
MonoATT: Online Monocular 3D Object Detection With Adaptive Token Transformer
Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, Minyi Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17493-17503
Mobile monocular 3D object detection (Mono3D) (e.g., on a vehicle, a drone, or a robot) is an important yet challenging task. Existing transformer-based offline Mono3D models adopt grid-based vision tokens, which is suboptimal when using coarse tokens due to the limited available computational power. In this paper, we propose an online Mono3D framework, called MonoATT, which leverages a novel vision transformer with heterogeneous tokens of varying shapes and sizes to facilitate mobile Mono3D. The core idea of MonoATT is to adaptively assign finer tokens to areas of more significance before utilizing a transformer to enhance Mono3D. To this end, we first use prior knowledge to design a scoring network for selecting the most important areas of the image, and then propose a token clustering and merging network with an attention mechanism to gradually merge tokens around the selected areas in multiple stages. Finally, a pixel-level feature map is reconstructed from heterogeneous tokens before employing a SOTA Mono3D detector as the underlying detection core. Experiment results on the real-world KITTI dataset demonstrate that MonoATT can effectively improve the Mono3D accuracy for both near and far objects and guarantee low latency. MonoATT yields the best performance compared with the state-of-the-art methods by a large margin and is ranked number one on the KITTI 3D benchmark.
****
Image Quality-Aware Diagnosis via Meta-Knowledge Co-Embedding
Haoxuan Che, Siyu Chen, Hao Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19819-19829
Medical images usually suffer from image degradation in clinical practice, leading to decreased performance of deep learning-based models. To resolve this problem, most previous works have focused on filtering out degradation-causing low-quality images while ignoring their potential value for models. Through effectively learning and leveraging the knowledge of degradations, models can better resist their adverse effects and avoid misdiagnosis. In this paper, we raise the problem of image quality-aware diagnosis, which aims to take advantage of low-quality images and image quality labels to achieve a more accurate and robust diagnosis. However, the diversity of degradations and superficially unrelated targets between image quality assessment and disease diagnosis makes it still quite challenging to effectively leverage quality labels to assist diagnosis. Thus, to tackle these issues, we propose a novel meta-knowledge co-embedding network, consisting of two subnets: Task Net and Meta Learner. Task Net constructs an explicit quality information utilization mechanism to enhance diagnosis via knowledge co-embedding features, while Meta Learner ensures the effectiveness and constrains the semantics of these features via meta-learning and joint-encoding masking. Superior performance on five datasets with four widely-used medical imaging modalities demonstrates the effectiveness and generalizability of our method.
****
A-Cap: Anticipation Captioning With Commonsense Knowledge
Duc Minh Vo, Quoc-An Luong, Akihiro Sugimoto, Hideki Nakayama; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10824-10833
Humans possess the capacity to reason about the future based on a sparse collection of visual cues acquired over time. In order to emulate this ability, we introduce a novel task called Anticipation Captioning, which generates a caption for an unseen oracle image using a sparsely temporally-ordered set of images. To tackle this new task, we propose a model called A-CAP, which incorporates commonsense knowledge into a pre-trained vision-language model, allowing it to anticipate the caption. Through both qualitative and quantitative evaluations on a customized visual storytelling dataset, A-CAP outperforms other image captioning methods and establishes a strong baseline for anticipation captioning. We also addres

s the challenges inherent in this task.
****
Learning 3D Representations From 2D Pre-Trained Models via Image-to-Point Masked Autoencoders
Renrui Zhang, Liuhui Wang, Yu Qiao, Peng Gao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21769-21780
Pre-training by numerous image data has become de-facto for robust 2D representations. In contrast, due to the expensive data processing, a paucity of 3D datasets severely hinders the learning for high-quality 3D features. In this paper, we propose an alternative to obtain superior 3D representations from 2D pre-trained models via Image-to-Point Masked Autoencoders, named as I2P-MAE. By self-supervised pre-training, we leverage the well learned 2D knowledge to guide 3D masked autoencoding, which reconstructs the masked point tokens with an encoder-decoder architecture. Specifically, we first utilize off-the-shelf 2D models to extract the multi-view visual features of the input point cloud, and then conduct two types of image-to-point learning schemes. For one, we introduce a 2D-guided masking strategy that maintains semantically important point tokens to be visible. Compared to random masking, the network can better concentrate on significant 3D structures with key spatial cues. For another, we enforce these visible tokens to reconstruct multi-view 2D features after the decoder. This enables the network to effectively inherit high-level 2D semantics for discriminative 3D modeling. Aided by our image-to-point pre-training, the frozen I2P-MAE, without any fine-tuning, achieves 93.4% accuracy for linear SVM on ModelNet40, competitive to existing fully trained methods. By further fine-tuning on on ScanObjectNN's hardest split, I2P-MAE attains the state-of-the-art 90.11% accuracy, +3.68% to the second-best, demonstrating superior transferable capacity. Code is available at https://github.com/ZrrSkywalker/I2P-MAE.
****
BEVFormer v2: Adapting Modern Image Backbones to Bird's-Eye-View Recognition via Perspective Supervision
Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17830-17839
We present a novel bird's-eye-view (BEV) detector with perspective supervision, which converges faster and better suits modern image backbones. Existing state-of-the-art BEV detectors are often tied to certain depth pre-trained backbones like VoVNet, hindering the synergy between booming image backbones and BEV detectors. To address this limitation, we prioritize easing the optimization of BEV detectors by introducing perspective space supervision. To this end, we propose a two-stage BEV detector, where proposals from the perspective head are fed into the bird's-eye-view head for final predictions. To evaluate the effectiveness of our model, we conduct extensive ablation studies focusing on the form of supervision and the generality of the proposed detector. The proposed method is verified with a wide spectrum of traditional and modern image backbones and achieves new SoTA results on the large-scale nuScenes dataset. The code shall be released soon.
****
Object Discovery From Motion-Guided Tokens
Zhipeng Bao, Pavel Tokmakov, Yu-Xiong Wang, Adrien Gaidon, Martial Hebert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22972-22981
Object discovery -- separating objects from the background without manual labels -- is a fundamental open challenge in computer vision. Previous methods struggle to go beyond clustering of low-level cues, whether handcrafted (e.g., color, texture) or learned (e.g., from auto-encoders). In this work, we augment the auto-encoder representation learning framework with two key components: motion-guidance and mid-level feature tokenization. Although both have been separately investigated, we introduce a new transformer decoder showing that their benefits can

compound thanks to motion-guided vector quantization. We show that our architecture effectively leverages the synergy between motion and tokenization, improving upon the state of the art on both synthetic and real datasets. Our approach enables the emergence of interpretable object-specific mid-level features, demonstrating the benefits of motion-guidance (no labeling) and quantization (interpretability, memory efficiency).
****
Domain Generalized Stereo Matching via Hierarchical Visual Transformation
Tianyu Chang, Xun Yang, Tianzhu Zhang, Meng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9559-9568
Recently, deep Stereo Matching (SM) networks have shown impressive performance and attracted increasing attention in computer vision. However, existing deep SM networks are prone to learn dataset-dependent shortcuts, which fail to generalize well on unseen realistic datasets. This paper takes a step towards training robust models for the domain generalized SM task, which mainly focuses on learning shortcut-invariant representation from synthetic data to alleviate the domain shifts. Specifically, we propose a Hierarchical Visual Transformation (HVT) network to 1) first transform the training sample hierarchically into new domains with diverse distributions from three levels: Global, Local, and Pixel, 2) then maximize the visual discrepancy between the source domain and new domains, and minimize the cross-domain feature inconsistency to capture domain-invariant features. In this way, we can prevent the model from exploiting the artifacts of synthetic stereo images as shortcut features, thereby estimating the disparity maps more effectively based on the learned robust and shortcut-invariant representation. We integrate our proposed HVT network with SOTA SM networks and evaluate its effectiveness on several public SM benchmark datasets. Extensive experiments clearly show that the HVT network can substantially enhance the performance of existing SM networks in synthetic-to-realistic domain generalization.
****
Deep Semi-Supervised Metric Learning With Mixed Label Propagation
Furen Zhuang, Pierre Moulin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3429-3438
Metric learning requires the identification of far-apart similar pairs and close dissimilar pairs during training, and this is difficult to achieve with unlabeled data because pairs are typically assumed to be similar if they are close. We present a novel metric learning method which circumvents this issue by identifying hard negative pairs as those which obtain dissimilar labels via label propagation (LP), when the edge linking the pair of data is removed in the affinity matrix. In so doing, the negative pairs can be identified despite their proximity, and we are able to utilize this information to significantly improve LP's ability to identify far-apart positive pairs and close negative pairs. This results in a considerable improvement in semi-supervised metric learning performance as evidenced by recall, precision and Normalized Mutual Information (NMI) performance metrics on Content-based Information Retrieval (CBIR) applications.
****
Adapting Shortcut With Normalizing Flow: An Efficient Tuning Framework for Visual Recognition
Yaoming Wang, Bowen Shi, Xiaopeng Zhang, Jin Li, Yuchen Liu, Wenrui Dai, Chenglin Li, Hongkai Xiong, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15965-15974
Pretraining followed by fine-tuning has proven to be effective in visual recognition tasks. However, fine-tuning all parameters can be computationally expensive, particularly for large-scale models. To mitigate the computational and storage demands, recent research has explored Parameter-Efficient Fine-Tuning (PEFT), which focuses on tuning a minimal number of parameters for efficient adaptation. Existing methods, however, fail to analyze the impact of the additional parameters on the model, resulting in an unclear and suboptimal tuning process. In this paper, we introduce a novel and effective PEFT paradigm, named SNF (Shortcut adaptation via Normalization Flow), which utilizes normalizing flows to adjust the shortcut layers. We highlight that layers without Lipschitz constraints can lead

to error propagation when adapting to downstream datasets. Since modifying the over-parameterized residual connections in these layers is expensive, we focus on adjusting the cheap yet crucial shortcuts. Moreover, learning new information with few parameters in PEFT can be challenging, and information loss can result in label information degradation. To address this issue, we propose an information-preserving normalizing flow. Experimental results demonstrate the effectiveness of SNF. Specifically, with only 0.036M parameters, SNF surpasses previous approaches on both the FGVC and VTAB-1k benchmarks using ViT/B-16 as the backbone. The code is available at https://github.com/Wang-Yaoming/SNF
****
Unpaired Image-to-Image Translation With Shortest Path Regularization
Shaoan Xie, Yanwu Xu, Mingming Gong, Kun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10177-10187
Unpaired image-to-image translation aims to learn proper mappings that can map images from one domain to another domain while preserving the content of the input image. However, with large enough capacities, the network can learn to map the inputs to any random permutation of images in another domain. Existing methods treat two domains as discrete and propose different assumptions to address this problem. In this paper, we start from a different perspective and consider the paths connecting the two domains. We assume that the optimal path length between the input and output image should be the shortest among all possible paths. Based on this assumption, we propose a new method to allow generating images along the path and present a simple way to encourage the network to find the shortest path without pair information. Extensive experiments on various tasks demonstrate the superiority of our approach.
****
MotionDiffuser: Controllable Multi-Agent Motion Prediction Using Diffusion
Chiyu “Max” Jiang, Andre Cornman, Cheolho Park, Benjamin Sapp, Yin Zhou, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9644-9653
We present MotionDiffuser, a diffusion based representation for the joint distribution of future trajectories over multiple agents. Such representation has several key advantages: first, our model learns a highly multimodal distribution that captures diverse future outcomes. Second, the simple predictor design requires only a single L2 loss training objective, and does not depend on trajectory anchors. Third, our model is capable of learning the joint distribution for the motion of multiple agents in a permutation-invariant manner. Furthermore, we utilize a compressed trajectory representation via PCA, which improves model performance and allows for efficient computation of the exact sample log probability. Subsequently, we propose a general constrained sampling framework that enables controlled trajectory sampling based on differentiable cost functions. This strategy enables a host of applications such as enforcing rules and physical priors, or creating tailored simulation scenarios. MotionDiffuser can be combined with existing backbone architectures to achieve top motion forecasting results. We obtain state-of-the-art results for multi-agent motion prediction on the Waymo Open Motion Dataset.
****
OVTrack: Open-Vocabulary Multiple Object Tracking
Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5567-5577
The ability to recognize, localize and track dynamic objects in a scene is fundamental to many real-world applications, such as self-driving and robotic systems. Yet, traditional multiple object tracking (MOT) benchmarks rely only on a few object categories that hardly represent the multitude of possible objects that are encountered in the real world. This leaves contemporary MOT methods limited to a small set of pre-defined object categories. In this paper, we address this limitation by tackling a novel task, open-vocabulary MOT, that aims to evaluate tracking beyond pre-defined training categories. We further develop OVTrack, an open-vocabulary tracker that is capable of tracking arbitrary object classes. Its

design is based on two key ingredients: First, leveraging vision-language model
s for both classification and association via knowledge distillation; second, a
data hallucination strategy for robust appearance feature learning from denoisin
g diffusion probabilistic models. The result is an extremely data-efficient open
-vocabulary tracker that sets a new state-of-the-art on the large-scale, large-v
ocabulary TAO benchmark, while being trained solely on static images. The projec
t page is at https://www.vis.xyz/pub/ovtrack/.
****
ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders
Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweo
n, Saining Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2023, pp. 16133-16142
Driven by improved architectures and better representation learning frameworks,
the field of visual recognition has enjoyed rapid modernization and performance
boost in the early 2020s. For example, modern ConvNets, represented by ConvNeXt
models, have demonstrated strong performance across different application scenar
ios. Like many other architectures, ConvNeXt models were designed under the supe
rvised learning setting with ImageNet labels. It is natural to expect ConvNeXt c
an also benefit from state-of-the-art self-supervised learning frameworks such a
s masked autoencoders (MAE), which was originally designed with Transformers. Ho
wever, we show that simply combining the two designs yields subpar performance.
In this paper, we develop an efficient and fully-convolutional masked autoencode
r framework. We then upgrade the ConvNeXt architecture with a new Global Respons
e Normalization (GRN) layer. GRN enhances inter-channel feature competition and
is crucial for pre-training with masked input. The new model family, dubbed Conv
NeXt V2, is a complete training recipe that synergizes both the architectural im
provement and the advancement in self-supervised learning. With ConvNeXt V2, we
are able to significantly advance pure ConvNets' performance across different re
cognition benchmarks including ImageNet classification, ADE20K segmentation and
COCO detection. To accommodate different use cases, we provide pre-trained ConvN
eXt V2 models of a wide range of complexity: from an efficient 3.7M-parameter At
to model that achieves 76.8% top-1 accuracy on ImageNet, to a 650M Huge model th
at can reach a state-of-the-art 88.9% accuracy using public training data only.
****
Hyperspherical Embedding for Point Cloud Completion
Junming Zhang, Haomeng Zhang, Ram Vasudevan, Matthew Johnson-Roberson; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2023, pp. 5323-5332
Most real-world 3D measurements from depth sensors are incomplete, and to addres
s this issue the point cloud completion task aims to predict the complete shapes
 of objects from partial observations. Previous works often adapt an encoder-dec
oder architecture, where the encoder is trained to extract embeddings that are u
sed as inputs to generate predictions from the decoder. However, the learned emb
eddings have sparse distribution in the feature space, which leads to worse gene
ralization results during testing. To address these problems, this paper propose
s a hyperspherical module, which transforms and normalizes embeddings from the e
ncoder to be on a unit hypersphere. With the proposed module, the magnitude and
direction of the output hyperspherical embedding are decoupled and only the dire
ctional information is optimized. We theoretically analyze the hyperspherical em
bedding and show that it enables more stable training with a wider range of lear
ning rates and more compact embedding distributions. Experiment results show con
sistent improvement of point cloud completion in both single-task and multi-task
 learning, which demonstrates the effectiveness of the proposed method.
****
Event-Based Video Frame Interpolation With Cross-Modal Asymmetric Bidirectional
Motion Fields
Taewoo Kim, Yujeong Chae, Hyun-Kurl Jang, Kuk-Jin Yoon; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1803
2-18042
Video Frame Interpolation (VFI) aims to generate intermediate video frames betwe

en consecutive input frames. Since the event cameras are bio-inspired sensors th
at only encode brightness changes with a micro-second temporal resolution, sever
al works utilized the event camera to enhance the performance of VFI. However, e
xisting methods estimate bidirectional inter-frame motion fields with only event
s or approximations, which can not consider the complex motion in real-world sce
narios. In this paper, we propose a novel event-based VFI framework with cross-m
odal asymmetric bidirectional motion field estimation. In detail, our EIF-BiOFNe
t utilizes each valuable characteristic of the events and images for direct esti
mation of inter-frame motion fields without any approximation methods.Moreover,
we develop an interactive attention-based frame synthesis network to efficiently
 leverage the complementary warping-based and synthesis-based features. Finally,
 we build a large-scale event-based VFI dataset, ERF-X170FPS, with a high frame
rate, extreme motion, and dynamic textures to overcome the limitations of previo
us event-based VFI datasets. Extensive experimental results validate that our me
thod shows significant performance improvement over the state-of-the-art VFI met
hods on various datasets.Our project pages are available at: https://github.com/
intelpro/CBMNet
****
Unsupervised Deep Asymmetric Stereo Matching With Spatially-Adaptive Self-Simila
rity
Taeyong Song, Sunok Kim, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13672-13680
Unsupervised stereo matching has received a lot of attention since it enables th
e learning of disparity estimation without ground-truth data. However, most of t
he unsupervised stereo matching algorithms assume that the left and right images
 have consistent visual properties, i.e., symmetric, and easily fail when the st
ereo images are asymmetric. In this paper, we present a novel spatially-adaptive
 self-similarity (SASS) for unsupervised asymmetric stereo matching. It extends
the concept of self-similarity and generates deep features that are robust to th
e asymmetries. The sampling patterns to calculate self-similarities are adaptive
ly generated throughout the image regions to effectively encode diverse patterns
. In order to learn the effective sampling patterns, we design a contrastive sim
ilarity loss with positive and negative weights. Consequently, SASS is further e
ncouraged to encode asymmetry-agnostic features, while maintaining the distincti
veness for stereo correspondence. We present extensive experimental results incl
uding ablation studies and comparisons with different methods, demonstrating eff
ectiveness of the proposed method under resolution and noise asymmetries.
****
QuantArt: Quantizing Image Style Transfer Towards High Visual Fidelity
Siyu Huang, Jie An, Donglai Wei, Jiebo Luo, Hanspeter Pfister; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 5947-5956
The mechanism of existing style transfer algorithms is by minimizing a hybrid lo
ss function to push the generated image toward high similarities in both content
 and style. However, this type of approach cannot guarantee visual fidelity, i.e
., the generated artworks should be indistinguishable from real ones. In this pa
per, we devise a new style transfer framework called QuantArt for high visual-fi
delity stylization. QuantArt pushes the latent representation of the generated a
rtwork toward the centroids of the real artwork distribution with vector quantiz
ation. By fusing the quantized and continuous latent representations, QuantArt a
llows flexible control over the generated artworks in terms of content preservat
ion, style similarity, and visual fidelity. Experiments on various style transfe
r settings show that our QuantArt framework achieves significantly higher visual
 fidelity compared with the existing style transfer methods.
****
TWINS: A Fine-Tuning Framework for Improved Transferability of Adversarial Robus
tness and Generalization
Ziquan Liu, Yi Xu, Xiangyang Ji, Antoni B. Chan; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16436-16446
Recent years have seen the ever-increasing importance of pre-trained models and

their downstream training in deep learning research and applications. At the same time, the defense for adversarial examples has been mainly investigated in the context of training from random initialization on simple classification tasks. To better exploit the potential of pre-trained models in adversarial robustness, this paper focuses on the fine-tuning of an adversarially pre-trained model in various classification tasks. Existing research has shown that since the robust pre-trained model has already learned a robust feature extractor, the crucial question is how to maintain the robustness in the pre-trained model when learning the downstream task. We study the model-based and data-based approaches for this goal and find that the two common approaches cannot achieve the objective of improving both generalization and adversarial robustness. Thus, we propose a novel statistics-based approach, Two-WIng NormliSation (TWINS) fine-tuning framework, which consists of two neural networks where one of them keeps the population means and variances of pre-training data in the batch normalization layers. Besides the robust information transfer, TWINS increases the effective learning rate without hurting the training stability since the relationship between a weight norm and its gradient norm in standard batch normalization layer is broken, resulting in a faster escape from the sub-optimal initialization and alleviating the robust overfitting. Finally, TWINS is shown to be effective on a wide range of image classification datasets in terms of both generalization and robustness.
****
VolRecon: Volume Rendering of Signed Ray Distance Functions for Generalizable Multi-View Reconstruction
Yufan Ren, Fangjinhua Wang, Tong Zhang, Marc Pollefeys, Sabine Süsstrunk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16685-16695
The success of the Neural Radiance Fields (NeRF) in novel view synthesis has inspired researchers to propose neural implicit scene reconstruction. However, most existing neural implicit reconstruction methods optimize per-scene parameters and therefore lack generalizability to new scenes. We introduce VolRecon, a novel generalizable implicit reconstruction method with Signed Ray Distance Function (SRDF). To reconstruct the scene with fine details and little noise, VolRecon combines projection features aggregated from multi-view features, and volume features interpolated from a coarse global feature volume. Using a ray transformer, we compute SRDF values of sampled points on a ray and then render color and depth. On DTU dataset, VolRecon outperforms SparseNeuS by about 30% in sparse view reconstruction and achieves comparable accuracy as MVSNet in full view reconstruction. Furthermore, our approach exhibits good generalization performance on the large-scale ETH3D benchmark.
****
Object-Aware Distillation Pyramid for Open-Vocabulary Object Detection
Luting Wang, Yi Liu, Penghui Du, Zihan Ding, Yue Liao, Qiaosong Qi, Biaolong Chen, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11186-11196
Open-vocabulary object detection aims to provide object detectors trained on a fixed set of object categories with the generalizability to detect objects described by arbitrary text queries. Previous methods adopt knowledge distillation to extract knowledge from Pretrained Vision-and-Language Models (PVLMs) and transfer it to detectors. However, due to the non-adaptive proposal cropping and single-level feature mimicking processes, they suffer from information destruction during knowledge extraction and inefficient knowledge transfer. To remedy these limitations, we propose an Object-Aware Distillation Pyramid (OADP) framework, including an Object-Aware Knowledge Extraction (OAKE) module and a Distillation Pyramid (DP) mechanism. When extracting object knowledge from PVLMs, the former adaptively transforms object proposals and adopts object-aware mask attention to obtain precise and complete knowledge of objects. The latter introduces global and block distillation for more comprehensive knowledge transfer to compensate for the missing relation information in object distillation. Extensive experiments show that our method achieves significant improvement compared to current methods. Especially on the MS-COCO dataset, our OADP framework reaches 35.6 mAP^N_50, su

rpassing the current state-of-the-art method by 3.3 mAP^N_50. Code is anonymousl
y provided in the supplementary materials.
****
Evolved Part Masking for Self-Supervised Learning
Zhanzhou Feng, Shiliang Zhang; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2023, pp. 10386-10395
Existing Masked Image Modeling methods apply fixed mask patterns to guide the se
lf-supervised training. As those patterns resort to different criteria to mask l
ocal regions, sticking to a fixed pattern leads to limited vision cues modeling
capability. This paper proposes an evolved part-based masking to pursue more gen
eral visual cues modeling in self-supervised learning. Our method is based on an
 adaptive part partition module, which leverages the vision model being trained
to construct a part graph, and partitions parts with graph cut. The accuracy of
partitioned parts is on par with the capability of the pre-trained model, leadin
g to evolved mask patterns at different training stages. It generates simple pat
terns at the initial training stage to learn low-level visual cues, which hence
evolves to eliminate accurate object parts to reinforce the learning of object s
emantics and contexts. Our method does not require extra pre-trained models or a
nnotations, and effectively ensures the training efficiency by evolving the trai
ning difficulty. Experiment results show that it substantially boosts the perfor
mance on various tasks including image classification, object detection, and sem
antic segmentation. For example, it outperforms the recent MAE by 0.69% on image
Net-1K classification and 1.61% on ADE20K segmentation with the same training ep
ochs.
****
MV-JAR: Masked Voxel Jigsaw and Reconstruction for LiDAR-Based Self-Supervised P
re-Training
Runsen Xu, Tai Wang, Wenwei Zhang, Runjian Chen, Jinkun Cao, Jiangmiao Pang, Dah
ua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2023, pp. 13445-13454
This paper introduces the Masked Voxel Jigsaw and Reconstruction (MV-JAR) method
 for LiDAR-based self-supervised pre-training and a carefully designed data-effi
cient 3D object detection benchmark on the Waymo dataset. Inspired by the scene-
voxel-point hierarchy in downstream 3D object detectors, we design masking and r
econstruction strategies accounting for voxel distributions in the scene and loc
al point distributions within the voxel. We employ a Reversed-Furthest-Voxel-Sam
pling strategy to address the uneven distribution of LiDAR points and propose MV
-JAR, which combines two techniques for modeling the aforementioned distribution
s, resulting in superior performance. Our experiments reveal limitations in prev
ious data-efficient experiments, which uniformly sample fine-tuning splits with
varying data proportions from each LiDAR sequence, leading to similar data diver
sity across splits. To address this, we propose a new benchmark that samples sce
ne sequences for diverse fine-tuning splits, ensuring adequate model convergence
 and providing a more accurate evaluation of pre-training methods. Experiments o
n our Waymo benchmark and the KITTI dataset demonstrate that MV-JAR consistently
 and significantly improves 3D detection performance across various data scales,
 achieving up to a 6.3% increase in mAPH compared to training from scratch. Code
s and the benchmark are available at https://github.com/SmartBot-PJLab/MV-JAR.
****
SlowLiDAR: Increasing the Latency of LiDAR-Based Detection Using Adversarial Exa
mples
Han Liu, Yuhao Wu, Zhiyuan Yu, Yevgeniy Vorobeychik, Ning Zhang; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
 pp. 5146-5155
LiDAR-based perception is a central component of autonomous driving, playing a k
ey role in tasks such as vehicle localization and obstacle detection. Since the
safety of LiDAR-based perceptual pipelines is critical to safe autonomous drivin
g, a number of past efforts have investigated its vulnerability under adversaria
l perturbations of raw point cloud inputs. However, most such efforts have focus
ed on investigating the impact of such perturbations on predictions (integrity),

and little has been done to understand the impact on latency (availability), a critical concern for real-time cyber-physical systems. We present the first systematic investigation of the availability of LiDAR detection pipelines, and SlowLiDAR, an adversarial perturbation attack that maximizes LiDAR detection runtime. The attack overcomes the technical challenges posed by the non-differentiable parts of the LiDAR detection pipelines by using differentiable proxies and uses a novel loss function that effectively captures the impact of adversarial perturbations on the execution time of the pipeline. Extensive experimental results show that SlowLiDAR can significantly increase the latency of the six most popular LiDAR detection pipelines while maintaining imperceptibility.
****
Learning a Sparse Transformer Network for Effective Image Deraining
Xiang Chen, Hao Li, Mingqiang Li, Jinshan Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5896-5905
Transformers-based methods have achieved significant performance in image deraining as they can model the non-local information which is vital for high-quality image reconstruction. In this paper, we find that most existing Transformers usually use all similarities of the tokens from the query-key pairs for the feature aggregation. However, if the tokens from the query are different from those of the key, the self-attention values estimated from these tokens also involve in feature aggregation, which accordingly interferes with the clear image restoration. To overcome this problem, we propose an effective DeRaining network, Sparse Transformer (DRSformer) that can adaptively keep the most useful self-attention values for feature aggregation so that the aggregated features better facilitate high-quality image reconstruction. Specifically, we develop a learnable top-k selection operator to adaptively retain the most crucial attention scores from the keys for each query for better feature aggregation. Simultaneously, as the naive feed-forward network in Transformers does not model the multi-scale information that is important for latent clear image restoration, we develop an effective mixed-scale feed-forward network to generate better features for image deraining. To learn an enriched set of hybrid features, which combines local context from CNN operators, we equip our model with mixture of experts feature compensator to present a cooperation refinement deraining scheme. Extensive experimental results on the commonly used benchmarks demonstrate that the proposed method achieves favorable performance against state-of-the-art approaches. The source code and trained models are available at https://github.com/cschenxiang/DRSformer.
****
Open-Set Semantic Segmentation for Point Clouds via Adversarial Prototype Framework
Jianan Li, Qiulei Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9425-9434
Recently, point cloud semantic segmentation has attracted much attention in computer vision. Most of the existing works in literature assume that the training and testing point clouds have the same object classes, but they are generally invalid in many real-world scenarios for identifying the 3D objects whose classes are not seen in the training set. To address this problem, we propose an Adversarial Prototype Framework (APF) for handling the open-set 3D semantic segmentation task, which aims to identify 3D unseen-class points while maintaining the segmentation performance on seen-class points. The proposed APF consists of a feature extraction module for extracting point features, a prototypical constraint module, and a feature adversarial module. The prototypical constraint module is designed to learn prototypes for each seen class from point features. The feature adversarial module utilizes generative adversarial networks to estimate the distribution of unseen-class features implicitly, and the synthetic unseen-class features are utilized to prompt the model to learn more effective point features and prototypes for discriminating unseen-class samples from the seen-class ones. Experimental results on two public datasets demonstrate that the proposed APF outperforms the comparative methods by a large margin in most cases.
****
CutMIB: Boosting Light Field Super-Resolution via Multi-View Image Blending

Zeyu Xiao, Yutong Liu, Ruisheng Gao, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1672-1682
Data augmentation (DA) is an efficient strategy for improving the performance of deep neural networks. Recent DA strategies have demonstrated utility in single image super-resolution (SR). Little research has, however, focused on the DA strategy for light field SR, in which multi-view information utilization is required. For the first time in light field SR, we propose a potent DA strategy called CutMIB to improve the performance of existing light field SR networks while keeping their structures unchanged. Specifically, CutMIB first cuts low-resolution (LR) patches from each view at the same location. Then CutMIB blends all LR patches to generate the blended patch and finally pastes the blended patch to the corresponding regions of high-resolution light field views, and vice versa. By doing so, CutMIB enables light field SR networks to learn from implicit geometric information during the training stage. Experimental results demonstrate that CutMIB can improve the reconstruction performance and the angular consistency of existing light field SR networks. We further verify the effectiveness of CutMIB on real-world light field SR and light field denoising. The implementation code is available at https://github.com/zeyuxiao1997/CutMIB.
****
Learning Attention As Disentangler for Compositional Zero-Shot Learning
Shaozhe Hao, Kai Han, Kwan-Yee K. Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15315-15324
Compositional zero-shot learning (CZSL) aims at learning visual concepts (i.e., attributes and objects) from seen compositions and combining concept knowledge into unseen compositions. The key to CZSL is learning the disentanglement of the attribute-object composition. To this end, we propose to exploit cross-attentions as compositional disentanglers to learn disentangled concept embeddings. For example, if we want to recognize an unseen composition "yellow flower", we can learn the attribute concept "yellow" and object concept "flower" from different yellow objects and different flowers respectively. To further constrain the disentanglers to learn the concept of interest, we employ a regularization at the attention level. Specifically, we adapt the earth mover's distance (EMD) as a feature similarity metric in the cross-attention module. Moreover, benefiting from concept disentanglement, we improve the inference process and tune the prediction score by combining multiple concept probabilities. Comprehensive experiments on three CZSL benchmark datasets demonstrate that our method significantly outperforms previous works in both closed- and open-world settings, establishing a new state-of-the-art. Project page: https://haoosz.github.io/ade-czsl/
****
DA-DETR: Domain Adaptive Detection Transformer With Information Fusion
Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, Xiaoqin Zhang, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23787-23798
The recent detection transformer (DETR) simplifies the object detection pipeline by removing hand-crafted designs and hyperparameters as employed in conventional two-stage object detectors. However, how to leverage the simple yet effective DETR architecture in domain adaptive object detection is largely neglected. Inspired by the unique DETR attention mechanisms, we design DA-DETR, a domain adaptive object detection transformer that introduces information fusion for effective transfer from a labeled source domain to an unlabeled target domain. DA-DETR introduces a novel CNN-Transformer Blender (CTBlender) that fuses the CNN features and Transformer features ingeniously for effective feature alignment and knowledge transfer across domains. Specifically, CTBlender employs the Transformer features to modulate the CNN features across multiple scales where the high-level semantic information and the low-level spatial information are fused for accurate object identification and localization. Extensive experiments show that DA-DETR achieves superior detection performance consistently across multiple widely adopted domain adaptation benchmarks.
****
Energy-Efficient Adaptive 3D Sensing

Brevin Tilmon, Zhanghao Sun, Sanjeev J. Koppal, Yicheng Wu, Georgios Evangelidis, Ramzi Zahreddine, Gurunandan Krishnan, Sizhuo Ma, Jian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5054-5063

Active depth sensing achieves robust depth estimation but is usually limited by the sensing range. Naively increasing the optical power can improve sensing range but induces eye-safety concerns for many applications, including autonomous robots and augmented reality. In this paper, we propose an adaptive active depth sensor that jointly optimizes range, power consumption, and eye-safety. The main observation is that we need not project light patterns to the entire scene but only to small regions of interest where depth is necessary for the application and passive stereo depth estimation fails. We theoretically compare this adaptive sensing scheme with other sensing strategies, such as full-frame projection, line scanning, and point scanning. We show that, to achieve the same maximum sensing distance, the proposed method consumes the least power while having the shortest (best) eye-safety distance. We implement this adaptive sensing scheme with two hardware prototypes, one with a phase-only spatial light modulator (SLM) and the other with a micro-electro-mechanical (MEMS) mirror and diffractive optical elements (DOE). Experimental results validate the advantage of our method and demonstrate its capability of acquiring higher quality geometry adaptively.

****

CR-FIQA: Face Image Quality Assessment by Learning Sample Relative Classifiability

Fadi Boutros, Meiling Fang, Marcel Klemt, Biying Fu, Naser Damer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5836-5845

Face image quality assessment (FIQA) estimates the utility of the captured image in achieving reliable and accurate recognition performance. This work proposes a novel FIQA method, CR-FIQA, that estimates the face image quality of a sample by learning to predict its relative classifiability. This classifiability is measured based on the allocation of the training sample feature representation in angular space with respect to its class center and the nearest negative class center. We experimentally illustrate the correlation between the face image quality and the sample relative classifiability. As such property is only observable for the training dataset, we propose to learn this property by probing internal network observations during the training process and utilizing it to predict the quality of unseen samples. Through extensive evaluation experiments on eight benchmarks and four face recognition models, we demonstrate the superiority of our proposed CR-FIQA over state-of-the-art (SOTA) FIQA algorithms.

****

Endpoints Weight Fusion for Class Incremental Semantic Segmentation

Jia-Wen Xiao, Chang-Bin Zhang, Jiekang Feng, Xialei Liu, Joost van de Weijer, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7204-7213

Class incremental semantic segmentation (CISS) focuses on alleviating catastrophic forgetting to improve discrimination. Previous work mainly exploit regularization (e.g., knowledge distillation) to maintain previous knowledge in the current model. However, distillation alone often yields limited gain to the model since only the representations of old and new models are restricted to be consistent. In this paper, we propose a simple yet effective method to obtain a model with strong memory of old knowledge, named Endpoints Weight Fusion (EWF). In our method, the model containing old knowledge is fused with the model retaining new knowledge in a dynamic fusion manner, strengthening the memory of old classes in ever-changing distributions. In addition, we analyze the relation between our fusion strategy and a popular moving average technique EMA, which reveals why our method is more suitable for class-incremental learning. To facilitate parameter fusion with closer distance in the parameter space, we use distillation to enhance the optimization process. Furthermore, we conduct experiments on two widely used datasets, achieving the state-of-the-art performance.

****

GeneCIS: A Benchmark for General Conditional Image Similarity
Sagar Vaze, Nicolas Carion, Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6862-6872
We argue that there are many notions of 'similarity' and that models, like humans, should be able to adapt to these dynamically. This contrasts with most representation learning methods, supervised or self-supervised, which learn a fixed embedding function and hence implicitly assume a single notion of similarity. For instance, models trained on ImageNet are biased towards object categories, while a user might prefer the model to focus on colors, textures or specific elements in the scene. In this paper, we propose the GeneCIS ('genesis') benchmark, which measures models' ability to adapt to a range of similarity conditions. Extending prior work, our benchmark is designed for zero-shot evaluation only, and hence considers an open-set of similarity conditions. We find that baselines from powerful CLIP models struggle on GeneCIS and that performance on the benchmark is only weakly correlated with ImageNet accuracy, suggesting that simply scaling existing methods is not fruitful. We further propose a simple, scalable solution based on automatically mining information from existing image-caption datasets. We find our method offers a substantial boost over the baselines on GeneCIS, and further improves zero-shot performance on related image retrieval benchmarks. In fact, though evaluated zero-shot, our model surpasses state-of-the-art supervised models on MIT-States.
****
MetaViewer: Towards a Unified Multi-View Representation
Ren Wang, Haoliang Sun, Yuling Ma, Xiaoming Xi, Yilong Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11590-11599
Existing multi-view representation learning methods typically follow a specific-to-uniform pipeline, extracting latent features from each view and then fusing or aligning them to obtain the unified object representation. However, the manually pre-specified fusion functions and aligning criteria could potentially degrade the quality of the derived representation. To overcome them, we propose a novel uniform-to-specific multi-view learning framework from a meta-learning perspective, where the unified representation no longer involves manual manipulation but is automatically derived from a meta-learner named MetaViewer. Specifically, we formulated the extraction and fusion of view-specific latent features as a nested optimization problem and solved it by using a bi-level optimization scheme. In this way, MetaViewer automatically fuses view-specific features into a unified one and learns the optimal fusion scheme by observing reconstruction processes from the unified to the specific over all views. Extensive experimental results in downstream classification and clustering tasks demonstrate the efficiency and effectiveness of the proposed method.
****
MD-VQA: Multi-Dimensional Quality Assessment for UGC Live Videos
Zicheng Zhang, Wei Wu, Wei Sun, Danyang Tu, Wei Lu, Xiongkuo Min, Ying Chen, Guangtao Zhai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1746-1755
User-generated content (UGC) live videos are often bothered by various distortions during capture procedures and thus exhibit diverse visual qualities. Such source videos are further compressed and transcoded by media server providers before being distributed to end-users. Because of the flourishing of UGC live videos, effective video quality assessment (VQA) tools are needed to monitor and perceptually optimize live streaming videos in the distributing process. Unfortunately, existing compressed UGC VQA databases are either small in scale or employ high-quality UGC videos as source videos, so VQA models developed on these databases have limited abilities to evaluate UGC live videos. In this paper, we address UGC Live VQA problems by constructing a first-of-a-kind subjective UGC Live VQA database and developing an effective evaluation tool. Concretely, 418 source UGC videos are collected in real live streaming scenarios and 3,762 compressed ones at different bit rates are generated for the subsequent subjective VQA experiments. Based on the built database, we develop a Multi-Dimensional VQA (MD-VQA) eva

luator to measure the visual quality of UGC live videos from semantic, distortion, and motion aspects respectively. Extensive experimental results show that MD-VQA achieves state-of-the-art performance on both our UGC Live VQA database and existing compressed UGC VQA databases.
****
Vision Transformers Are Good Mask Auto-Labelers
Shiyi Lan, Xitong Yang, Zhiding Yu, Zuxuan Wu, Jose M. Alvarez, Anima Anandkumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23745-23755
We propose Mask Auto-Labeler (MAL), a high-quality Transformer-based mask auto-labeling framework for instance segmentation using only box annotations. MAL takes box-cropped images as inputs and conditionally generates their mask pseudo-labels.We show that Vision Transformers are good mask auto-labelers. Our method significantly reduces the gap between auto-labeling and human annotation regarding mask quality. Instance segmentation models trained using the MAL-generated masks can nearly match the performance of their fully-supervised counterparts, retaining up to 97.4% performance of fully supervised models. The best model achieves 44.1% mAP on COCO instance segmentation (test-dev 2017), outperforming state-of-the-art box-supervised methods by significant margins. Qualitative results indicate that masks produced by MAL are, in some cases, even better than human annotations.
****
Neural Transformation Fields for Arbitrary-Styled Font Generation
Bin Fu, Junjun He, Jianjun Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22438-22447
Few-shot font generation (FFG), aiming at generating font images with a few samples, is an emerging topic in recent years due to the academic and commercial values. Typically, the FFG approaches follow the style-content disentanglement paradigm, which transfers the target font styles to characters by combining the content representations of source characters and the style codes of reference samples. Most existing methods attempt to increase font generation ability via exploring powerful style representations, which may be a sub-optimal solution for the FFG task due to the lack of modeling spatial transformation in transferring font styles. In this paper, we model font generation as a continuous transformation process from the source character image to the target font image via the creation and dissipation of font pixels, and embed the corresponding transformations into a neural transformation field. With the estimated transformation path, the neural transformation field generates a set of intermediate transformation results via the sampling process, and a font rendering formula is developed to accumulate them into the target font image. Extensive experiments show that our method achieves state-of-the-art performance on few-shot font generation task, which demonstrates the effectiveness of our proposed model. Our implementation is available at: https://github.com/fubinfb/NTF.
****
Spring: A High-Resolution High-Detail Dataset and Benchmark for Scene Flow, Optical Flow and Stereo
Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, Andrés Bruhn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4981-4991
While recent methods for motion and stereo estimation recover an unprecedented amount of details, such highly detailed structures are neither adequately reflected in the data of existing benchmarks nor their evaluation methodology. Hence, we introduce Spring -- a large, high-resolution, high-detail, computer-generated benchmark for scene flow, optical flow, and stereo. Based on rendered scenes from the open-source Blender movie "Spring", it provides photo-realistic HD datasets with state-of-the-art visual effects and ground truth training data. Furthermore, we provide a website to upload, analyze and compare results. Using a novel evaluation methodology based on a super-resolved UHD ground truth, our Spring benchmark can assess the quality of fine structures and provides further detailed performance statistics on different image regions. Regarding the number of ground

truth frames, Spring is 60x larger than the only scene flow benchmark, KITTI 2015, and 15x larger than the well-established MPI Sintel optical flow benchmark. Initial results for recent methods on our benchmark show that estimating fine details is indeed challenging, as their accuracy leaves significant room for improvement. The Spring benchmark and the corresponding datasets are available at http://spring-benchmark.org.
****
EDICT: Exact Diffusion Inversion via Coupled Transformations
Bram Wallace, Akash Gokul, Nikhil Naik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22532-22541
Finding an initial noise vector that produces an input image when fed into the diffusion process (known as inversion) is an important problem in denoising diffusion models (DDMs), with applications for real image editing. The standard approach for real image editing with inversion uses denoising diffusion implicit models (DDIMs) to deterministically noise the image to the intermediate state along the path that the denoising would follow given the original conditioning. However, DDIM inversion for real images is unstable as it relies on local linearization assumptions, which result in the propagation of errors, leading to incorrect image reconstruction and loss of content. To alleviate these problems, we propose Exact Diffusion Inversion via Coupled Transformations (EDICT), an inversion method that draws inspiration from affine coupling layers. EDICT enables mathematically exact inversion of real and model-generated images by maintaining two coupled noise vectors which are used to invert each other in an alternating fashion. Using Stable Diffusion [25], a state-of-the-art latent diffusion model, we demonstrate that EDICT successfully reconstructs real images with high fidelity. On complex image datasets like MS-COCO, EDICT reconstruction significantly outperforms DDIM, improving the mean square error of reconstruction by a factor of two. Using noise vectors inverted from real images, EDICT enables a wide range of image edits--from local and global semantic edits to image stylization--while maintaining fidelity to the original image structure. EDICT requires no model training/finetuning, prompt tuning, or extra data and can be combined with any pretrained DDM.
****
Natural Language-Assisted Sign Language Recognition
Ronglai Zuo, Fangyun Wei, Brian Mak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14890-14900
Sign languages are visual languages which convey information by signers' handshape, facial expression, body movement, and so forth. Due to the inherent restriction of combinations of these visual ingredients, there exist a significant number of visually indistinguishable signs (VISigns) in sign languages, which limits the recognition capacity of vision neural networks. To mitigate the problem, we propose the Natural Language-Assisted Sign Language Recognition (NLA-SLR) framework, which exploits semantic information contained in glosses (sign labels). First, for VISigns with similar semantic meanings, we propose language-aware label smoothing by generating soft labels for each training sign whose smoothing weights are computed from the normalized semantic similarities among the glosses to ease training. Second, for VISigns with distinct semantic meanings, we present an inter-modality mixup technique which blends vision and gloss features to further maximize the separability of different signs under the supervision of blended labels. Besides, we also introduce a novel backbone, video-keypoint network, which not only models both RGB videos and human body keypoints but also derives knowledge from sign videos of different temporal receptive fields. Empirically, our method achieves state-of-the-art performance on three widely-adopted benchmarks: MSASL, WLASL, and NMFs-CSL. Codes are available at https://github.com/FangyunWei/SLRT.
****
MAESTER: Masked Autoencoder Guided Segmentation at Pixel Resolution for Accurate, Self-Supervised Subcellular Structure Recognition
Ronald Xie, Kuan Pang, Gary D. Bader, Bo Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3292-3301

Accurate segmentation of cellular images remains an elusive task due to the intrinsic variability in morphology of biological structures. Complete manual segmentation is unfeasible for large datasets, and while supervised methods have been proposed to automate segmentation, they often rely on manually generated ground truths which are especially challenging and time consuming to generate in biology due to the requirement of domain expertise. Furthermore, these methods have limited generalization capacity, requiring additional manual labels to be generated for each dataset and use case. We introduce MAESTER (Masked AutoEncoder guided SegmenTation at pixEl Resolution), a self-supervised method for accurate, subcellular structure segmentation at pixel resolution. MAESTER treats segmentation as a representation learning and clustering problem. Specifically, MAESTER learns semantically meaningful token representations of multi-pixel image patches while simultaneously maintaining a sufficiently large field of view for contextual learning. We also develop a cover-and-stride inference strategy to achieve pixel-level subcellular structure segmentation. We evaluated MAESTER on a publicly available volumetric electron microscopy (VEM) dataset of primary mouse pancreatic islets beta cells and achieved upwards of 29.1% improvement over state-of-the-art under the same evaluation criteria. Furthermore, our results are competitive against supervised methods trained on the same tasks, closing the gap between self-supervised and supervised approaches. MAESTER shows promise for alleviating the critical bottleneck of ground truth generation for imaging related data analysis and thereby greatly increasing the rate of biological discovery.
****
Learning Semantic Relationship Among Instances for Image-Text Matching
Zheren Fu, Zhendong Mao, Yan Song, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15159-15168
Image-text matching, a bridge connecting image and language, is an important task, which generally learns a holistic cross-modal embedding to achieve a high-quality semantic alignment between the two modalities. However, previous studies only focus on capturing fragment-level relation within a sample from a particular modality, e.g., salient regions in an image or text words in a sentence, where they usually pay less attention to capturing instance-level interactions among samples and modalities, e.g., multiple images and texts. In this paper, we argue that sample relations could help learn subtle differences for hard negative instances, and thus transfer shared knowledge for infrequent samples should be promising in obtaining better holistic embeddings. Therefore, we propose a novel hierarchical relation modeling framework (HREM), which explicitly capture both fragment- and instance-level relations to learn discriminative and robust cross-modal embeddings. Extensive experiments on Flickr30K and MS-COCO show our proposed method outperforms the state-of-the-art ones by 4%-10% in terms of rSum.
****
AeDet: Azimuth-Invariant Multi-View 3D Object Detection
Chengjian Feng, Zequn Jie, Yujie Zhong, Xiangxiang Chu, Lin Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21580-21588
Recent LSS-based multi-view 3D object detection has made tremendous progress, by processing the features in Brid-Eye-View (BEV) via the convolutional detector. However, the typical convolution ignores the radial symmetry of the BEV features and increases the difficulty of the detector optimization. To preserve the inherent property of the BEV features and ease the optimization, we propose an azimuth-equivariant convolution (AeConv) and an azimuth-equivariant anchor. The sampling grid of AeConv is always in the radial direction, thus it can learn azimuth-invariant BEV features. The proposed anchor enables the detection head to learn predicting azimuth-irrelevant targets. In addition, we introduce a camera-decoupled virtual depth to unify the depth prediction for the images with different camera intrinsic parameters. The resultant detector is dubbed Azimuth-equivariant Detector (AeDet). Extensive experiments are conducted on nuScenes, and AeDet achieves a 62.0% NDS, surpassing the recent multi-view 3D object detectors such as PETRv2 and BEVDepth by a large margin.

****
OCELOT: Overlapped Cell on Tissue Dataset for Histopathology
Jeongun Ryu, Aaron Valero Puche, JaeWoong Shin, Seonwook Park, Biagio Brattoli, Jinhee Lee, Wonkyung Jung, Soo Ick Cho, Kyunghyun Paeng, Chan-Young Ock, Donggeun Yoo, Sérgio Pereira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23902-23912
Cell detection is a fundamental task in computational pathology that can be used for extracting high-level medical information from whole-slide images. For accurate cell detection, pathologists often zoom out to understand the tissue-level structures and zoom in to classify cells based on their morphology and the surrounding context. However, there is a lack of efforts to reflect such behaviors by pathologists in the cell detection models, mainly due to the lack of datasets containing both cell and tissue annotations with overlapping regions. To overcome this limitation, we propose and publicly release OCELOT, a dataset purposely dedicated to the study of cell-tissue relationships for cell detection in histopathology. OCELOT provides overlapping cell and tissue annotations on images acquired from multiple organs. Within this setting, we also propose multi-task learning approaches that benefit from learning both cell and tissue tasks simultaneously. When compared against a model trained only for the cell detection task, our proposed approaches improve cell detection performance on 3 datasets: proposed OCELOT, public TIGER, and internal CARP datasets. On the OCELOT test set in particular, we show up to 6.79 improvement in F1-score. We believe the contributions of this paper, including the release of the OCELOT dataset at https://lunit-io.github.io/research/publications/ocelot are a crucial starting point toward the important research direction of incorporating cell-tissue relationships in computation pathology.
****
Global-to-Local Modeling for Video-Based 3D Human Pose and Shape Estimation
Xiaolong Shen, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8887-8896
Video-based 3D human pose and shape estimations are evaluated by intra-frame accuracy and inter-frame smoothness. Although these two metrics are responsible for different ranges of temporal consistency, existing state-of-the-art methods treat them as a unified problem and use monotonous modeling structures (e.g., RNN or attention-based block) to design their networks. However, using a single kind of modeling structure is difficult to balance the learning of short-term and long-term temporal correlations, and may bias the network to one of them, leading to undesirable predictions like global location shift, temporal inconsistency, and insufficient local details. To solve these problems, we propose to structurally decouple the modeling of long-term and short-term correlations in an end-to-end framework, Global-to-Local Transformer (GLoT). First, a global transformer is introduced with a Masked Pose and Shape Estimation strategy for long-term modeling. The strategy stimulates the global transformer to learn more inter-frame correlations by randomly masking the features of several frames. Second, a local transformer is responsible for exploiting local details on the human mesh and interacting with the global transformer by leveraging cross-attention. Moreover, a Hierarchical Spatial Correlation Regressor is further introduced to refine intra-frame estimations by decoupled global-local representation and implicit kinematic constraints. Our GLoT surpasses previous state-of-the-art methods with the lowest model parameters on popular benchmarks, i.e., 3DPW, MPI-INF-3DHP, and Human3.6M. Codes are available at https://github.com/sxl142/GLoT.
****
BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion
Michael J. Black, Priyanka Patel, Joachim Tesch, Jinlong Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8726-8737
We show, for the first time, that neural networks trained only on synthetic data achieve state-of-the-art accuracy on the problem of 3D human pose and shape (HP

S) estimation from real images. Previous synthetic datasets have been small, unrealistic, or lacked realistic clothing. Achieving sufficient realism is non-trivial and we show how to do this for full bodies in motion. Specifically, our BEDLAM dataset contains monocular RGB videos with ground-truth 3D bodies in SMPL-X format. It includes a diversity of body shapes, motions, skin tones, hair, and clothing. The clothing is realistically simulated on the moving bodies using commercial clothing physics simulation. We render varying numbers of people in realistic scenes with varied lighting and camera motions. We then train various HPS regressors using BEDLAM and achieve state-of-the-art accuracy on real-image benchmarks despite training with synthetic data. We use BEDLAM to gain insights into what model design choices are important for accuracy. With good synthetic training data, we find that a basic method like HMR approaches the accuracy of the current SOTA method (CLIFF). BEDLAM is useful for a variety of tasks and all images, ground truth bodies, 3D clothing, support code, and more are available for research purposes. Additionally, we provide detailed information about our synthetic data generation pipeline, enabling others to generate their own datasets. See the project page: https://bedlam.is.tue.mpg.de/.
****

Self-Supervised Image-to-Point Distillation via Semantically Tolerant Contrastive Loss

Anas Mahmoud, Jordan S. K. Hu, Tianshu Kuai, Ali Harakeh, Liam Paull, Steven L. Waslander; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7102-7110

An effective framework for learning 3D representations for perception tasks is distilling rich self-supervised image features via contrastive learning. However, image-to-point representation learning for autonomous driving datasets faces two main challenges: 1) the abundance of self-similarity, which results in the contrastive losses pushing away semantically similar point and image regions and thus disturbing the local semantic structure of the learned representations, and 2) severe class imbalance as pretraining gets dominated by over-represented classes. We propose to alleviate the self-similarity problem through a novel semantically tolerant image-to-point contrastive loss that takes into consideration the semantic distance between positive and negative image regions to minimize contrasting semantically similar point and image regions. Additionally, we address class imbalance by designing a class-agnostic balanced loss that approximates the degree of class imbalance through an aggregate sample-to-samples semantic similarity measure. We demonstrate that our semantically-tolerant contrastive loss with class balancing improves state-of-the-art 2D-to-3D representation learning in all evaluation settings on 3D semantic segmentation. Our method consistently outperforms state-of-the-art 2D-to-3D representation learning frameworks across a wide range of 2D self-supervised pretrained models.
****

ProtoCon: Pseudo-Label Refinement via Online Clustering and Prototypical Consistency for Efficient Semi-Supervised Learning

Islam Nassar, Munawar Hayat, Ehsan Abbasnejad, Hamid Rezatofighi, Gholamreza Haffari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11641-11650

Confidence-based pseudo-labeling is among the dominant approaches in semi-supervised learning (SSL). It relies on including high-confidence predictions made on unlabeled data as additional targets to train the model. We propose ProtoCon, a novel SSL method aimed at the less-explored label-scarce SSL where such methods usually underperform. ProtoCon refines the pseudo-labels by leveraging their nearest neighbours' information. The neighbours are identified as the training proceeds using an online clustering approach operating in an embedding space trained via a prototypical loss to encourage well-formed clusters. The online nature of ProtoCon allows it to utilise the label history of the entire dataset in one training cycle to refine labels in the following cycle without the need to store image embeddings. Hence, it can seamlessly scale to larger datasets at a low cost. Finally, ProtoCon addresses the poor training signal in the initial phase of training (due to fewer confident predictions) by introducing an auxiliary self-su

pervised loss. It delivers significant gains and faster convergence over state-of-the-art across 5 datasets, including CIFARs, ImageNet and DomainNet.
****
Image Super-Resolution Using T-Tetromino Pixels
Simon Grosche, Andy Regensky, Jürgen Seiler, André Kaup; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9989-9998
For modern high-resolution imaging sensors, pixel binning is performed in low-lighting conditions and in case high frame rates are required. To recover the original spatial resolution, single-image super-resolution techniques can be applied for upscaling. To achieve a higher image quality after upscaling, we propose a novel binning concept using tetromino-shaped pixels. It is embedded into the field of compressed sensing and the coherence is calculated to motivate the sensor layouts used. Next, we investigate the reconstruction quality using tetromino pixels for the first time in literature. Instead of using different types of tetrominoes as proposed elsewhere, we show that using a small repeating cell consisting of only four T-tetrominoes is sufficient. For reconstruction, we use a locally fully connected reconstruction (LFCR) network as well as two classical reconstruction methods from the field of compressed sensing. Using the LFCR network in combination with the proposed tetromino layout, we achieve superior image quality in terms of PSNR, SSIM, and visually compared to conventional single-image super-resolution using the very deep super-resolution (VDSR) network. For PSNR, a gain of up to +1.92 dB is achieved.
****
GFIE: A Dataset and Baseline for Gaze-Following From 2D to 3D in Indoor Environments
Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, Jingtai Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8907-8916
Gaze-following is a kind of research that requires locating where the person in the scene is looking automatically under the topic of gaze estimation. It is an important clue for understanding human intention, such as identifying objects or regions of interest to humans. However, a survey of datasets used for gaze-following tasks reveals defects in the way they collect gaze point labels. Manual labeling may introduce subjective bias and is labor-intensive, while automatic labeling with an eye-tracking device would alter the person's appearance. In this work, we introduce GFIE, a novel dataset recorded by a gaze data collection system we developed. The system is constructed with two devices, an Azure Kinect and a laser rangefinder, which generate the laser spot to steer the subject's attention as they perform in front of the camera. And an algorithm is developed to locate laser spots in images for annotating 2D/3D gaze targets and removing ground truth introduced by the spots. The whole procedure of collecting gaze behavior allows us to obtain unbiased labels in unconstrained environments semi-automatically. We also propose a baseline method with stereo field-of-view (FoV) perception for establishing a 2D/3D gaze-following benchmark on the GFIE dataset. Project page: https://sites.google.com/view/gfie.
****
Efficient Robust Principal Component Analysis via Block Krylov Iteration and CUR Decomposition
Shun Fang, Zhengqin Xu, Shiqian Wu, Shoulie Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1348-1357
Robust principal component analysis (RPCA) is widely studied in computer vision. Recently an adaptive rank estimate based RPCA has achieved top performance in low-level vision tasks without the prior rank, but both the rank estimate and RPCA optimization algorithm involve singular value decomposition, which requires extremely huge computational resource for large-scale matrices. To address these issues, an efficient RPCA (eRPCA) algorithm is proposed based on block Krylov iteration and CUR decomposition in this paper. Specifically, the Krylov iteration method is employed to approximate the eigenvalue decomposition in the rank estimation, which requires O(ndrq + n(rq)^2) for an (nxd) input matrix, in which q is

a parameter with a small value, r is the target rank. Based on the estimated ran k, CUR decomposition is adopted to replace SVD in updating low-rank matrix compo nent, whose complexity reduces from O(rnd) to O(r^2n) per iteration. Experimenta l results verify the efficiency and effectiveness of the proposed eRPCA over the state-of-the-art methods in various low-level vision applications.
****
VIVE3D: Viewpoint-Independent Video Editing Using 3D-Aware GANs
Anna Frühstück, Nikolaos Sarafianos, Yuanlu Xu, Peter Wonka, Tony Tung; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023, pp. 4446-4455
We introduce VIVE3D, a novel approach that extends the capabilities of image-bas ed 3D GANs to video editing and is able to represent the input video in an ident ity-preserving and temporally consistent way. We propose two new building blocks . First, we introduce a novel GAN inversion technique specifically tailored to 3 D GANs by jointly embedding multiple frames and optimizing for the camera parame ters. Second, besides traditional semantic face edits (e.g. for age and expressi on), we are the first to demonstrate edits that show novel views of the head ena bled by the inherent properties of 3D GANs and our optical flow-guided compositi ng technique to combine the head with the background video. Our experiments demo nstrate that VIVE3D generates high-fidelity face edits at consistent quality fro m a range of camera viewpoints which are composited with the original video in a temporally and spatially-consistent manner.
****
Unsupervised Sampling Promoting for Stochastic Human Trajectory Prediction
Guangyi Chen, Zhenhao Chen, Shunxing Fan, Kun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17874-1 7884
The indeterminate nature of human motion requires trajectory prediction systems to use a probabilistic model to formulate the multi-modality phenomenon and infe r a finite set of future trajectories. However, the inference processes of most existing methods rely on Monte Carlo random sampling, which is insufficient to c over the realistic paths with finite samples, due to the long tail effect of the predicted distribution. To promote the sampling process of stochastic predictio n, we propose a novel method, called BOsampler, to adaptively mine potential pat hs with Bayesian optimization in an unsupervised manner, as a sequential design strategy in which new prediction is dependent on the previously drawn samples. S pecifically, we model the trajectory sampling as a Gaussian process and construc t an acquisition function to measure the potential sampling value. This acquisit ion function applies the original distribution as prior and encourages exploring paths in the long-tail region. This sampling method can be integrated with exis ting stochastic predictive models without retraining. Experimental results on va rious baseline methods demonstrate the effectiveness of our method. The source c ode is released in this link.
****
BKinD-3D: Self-Supervised 3D Keypoint Discovery From Multi-View Videos
Jennifer J. Sun, Lili Karashchuk, Amil Dravid, Serim Ryou, Sonia Fereidooni, Joh n C. Tuthill, Aggelos Katsaggelos, Bingni W. Brunton, Georgia Gkioxari, Ann Kenn edy, Yisong Yue, Pietro Perona; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2023, pp. 9001-9010
Quantifying motion in 3D is important for studying the behavior of humans and ot her animals, but manual pose annotations are expensive and time-consuming to obt ain. Self-supervised keypoint discovery is a promising strategy for estimating 3 D poses without annotations. However, current keypoint discovery approaches comm only process single 2D views and do not operate in the 3D space. We propose a ne w method to perform self-supervised keypoint discovery in 3D from multi-view vid eos of behaving agents, without any keypoint or bounding box supervision in 2D o r 3D. Our method, BKinD-3D, uses an encoder-decoder architecture with a 3D volum etric heatmap, trained to reconstruct spatiotemporal differences across multiple views, in addition to joint length constraints on a learned 3D skeleton of the subject. In this way, we discover keypoints without requiring manual supervision

in videos of humans and rats, demonstrating the potential of 3D keypoint discovery for studying behavior.
****
StyleRF: Zero-Shot 3D Style Transfer of Neural Radiance Fields
Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8338-8348
3D style transfer aims to render stylized novel views of a 3D scene with multi-view consistency. However, most existing work suffers from a three-way dilemma over accurate geometry reconstruction, high-quality stylization, and being generalizable to arbitrary new styles. We propose StyleRF (Style Radiance Fields), an innovative 3D style transfer technique that resolves the three-way dilemma by performing style transformation within the feature space of a radiance field. StyleRF employs an explicit grid of high-level features to represent 3D scenes, with which high-fidelity geometry can be reliably restored via volume rendering. In addition, it transforms the grid features according to the reference style which directly leads to high-quality zero-shot style transfer. StyleRF consists of two innovative designs. The first is sampling-invariant content transformation that makes the transformation invariant to the holistic statistics of the sampled 3D points and accordingly ensures multi-view consistency. The second is deferred style transformation of 2D feature maps which is equivalent to the transformation of 3D points but greatly reduces memory footprint without degrading multi-view consistency. Extensive experiments show that StyleRF achieves superior 3D stylization quality with precise geometry reconstruction and it can generalize to various new styles in a zero-shot manner. Project website: https://kunhao-liu.github.io/StyleRF/
****
Semantic Prompt for Few-Shot Image Recognition
Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23581-23591
Few-shot learning is a challenging problem since only a few examples are provided to recognize a new class. Several recent studies exploit additional semantic information, e.g. text embeddings of class names, to address the issue of rare samples through combining semantic prototypes with visual prototypes. However, these methods still suffer from the spurious visual features learned from the rare support samples, resulting in limited benefits. In this paper, we propose a novel Semantic Prompt (SP) approach for few-shot learning. Instead of the naive exploitation of semantic information for remedying classifiers, we explore leveraging semantic information as prompts to tune the visual feature extraction network adaptively. Specifically, we design two complementary mechanisms to insert semantic prompts into the feature extractor: one is to enable the interaction between semantic prompts and patch embeddings along the spatial dimension via self-attention, another is to supplement visual features with the transformed semantic prompts along the channel dimension. By combining these two mechanisms, the feature extractor presents a better ability to attend to the class-specific features and obtains more generalized image representations with merely a few support samples. Through extensive experiments on four datasets, the proposed approach achieves promising results, improving the 1-shot learning accuracy by 3.67% on average.
****
Accidental Light Probes
Hong-Xing Yu, Samir Agarwala, Charles Herrmann, Richard Szeliski, Noah Snavely, Jiajun Wu, Deqing Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12521-12530
Recovering lighting in a scene from a single image is a fundamental problem in computer vision. While a mirror ball light probe can capture omnidirectional lighting, light probes are generally unavailable in everyday images. In this work, we study recovering lighting from accidental light probes (ALPs)---common, shiny objects like Coke cans, which often accidentally appear in daily scenes. We prop

ose a physically-based approach to model ALPs and estimate lighting from their appearances in single images. The main idea is to model the appearance of ALPs by photogrammetrically principled shading and to invert this process via differentiable rendering to recover incidental illumination. We demonstrate that we can put an ALP into a scene to allow high-fidelity lighting estimation. Our model can also recover lighting for existing images that happen to contain an ALP.
****

Iterative Vision-and-Language Navigation
Jacob Krantz, Shurjo Banerjee, Wang Zhu, Jason Corso, Peter Anderson, Stefan Lee, Jesse Thomason; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14921-14930

We present Iterative Vision-and-Language Navigation (IVLN), a paradigm for evaluating language-guided agents navigating in a persistent environment over time. Existing Vision-and-Language Navigation (VLN) benchmarks erase the agent's memory at the beginning of every episode, testing the ability to perform cold-start navigation with no prior information. However, deployed robots occupy the same environment for long periods of time. The IVLN paradigm addresses this disparity by training and evaluating VLN agents that maintain memory across tours of scenes that consist of up to 100 ordered instruction-following Room-to-Room (R2R) episodes, each defined by an individual language instruction and a target path. We present discrete and continuous Iterative Room-to-Room (IR2R) benchmarks comprising about 400 tours each in 80 indoor scenes. We find that extending the implicit memory of high-performing transformer VLN agents is not sufficient for IVLN, but agents that build maps can benefit from environment persistence, motivating a renewed focus on map-building agents in VLN.
****

DPE: Disentanglement of Pose and Expression for General Video Portrait Editing
Youxin Pang, Yong Zhang, Weize Quan, Yanbo Fan, Xiaodong Cun, Ying Shan, Dong-Ming Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 427-436

One-shot video-driven talking face generation aims at producing a synthetic talking video by transferring the facial motion from a video to an arbitrary portrait image. Head pose and facial expression are always entangled in facial motion and transferred simultaneously. However, the entanglement sets up a barrier for these methods to be used in video portrait editing directly, where it may require to modify the expression only while maintaining the pose unchanged. One challenge of decoupling pose and expression is the lack of paired data, such as the same pose but different expressions. Only a few methods attempt to tackle this challenge with the feat of 3D Morphable Models (3DMMs) for explicit disentanglement. But 3DMMs are not accurate enough to capture facial details due to the limited number of Blendshapes, which has side effects on motion transfer. In this paper, we introduce a novel self-supervised disentanglement framework to decouple pose and expression without 3DMMs and paired data, which consists of a motion editing module, a pose generator, and an expression generator. The editing module projects faces into a latent space where pose motion and expression motion can be disentangled, and the pose or expression transfer can be performed in the latent space conveniently via addition. The two generators render the modified latent codes to images, respectively. Moreover, to guarantee the disentanglement, we propose a bidirectional cyclic training strategy with well-designed constraints. Evaluations demonstrate our method can control pose or expression independently and be used for general video editing.
****

Adversarial Counterfactual Visual Explanations
Guillaume Jeanneret, Loïc Simon, Frédéric Jurie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16425-16435

Counterfactual explanations and adversarial attacks have a related goal: flipping output labels with minimal perturbations regardless of their characteristics. Yet, adversarial attacks cannot be used directly in a counterfactual explanation perspective, as such perturbations are perceived as noise and not as actionable and understandable image modifications. Building on the robust learning literat

ure, this paper proposes an elegant method to turn adversarial attacks into semantically meaningful perturbations, without modifying the classifiers to explain. The proposed approach hypothesizes that Denoising Diffusion Probabilistic Models are excellent regularizers for avoiding high-frequency and out-of-distribution perturbations when generating adversarial attacks. The paper's key idea is to build attacks through a diffusion model to polish them. This allows studying the target model regardless of its robustification level. Extensive experimentation shows the advantages of our counterfactual explanation approach over current State-of-the-Art in multiple testbeds.
****
MaLP: Manipulation Localization Using a Proactive Scheme
Vishal Asnani, Xi Yin, Tal Hassner, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12343-12352
Advancements in the generation quality of various Generative Models (GMs) has made it necessary to not only perform binary manipulation detection but also localize the modified pixels in an image. However, prior works termed as passive for manipulation localization exhibit poor generalization performance over unseen GMs and attribute modifications. To combat this issue, we propose a proactive scheme for manipulation localization, termed MaLP. We encrypt the real images by adding a learned template. If the image is manipulated by any GM, this added protection from the template not only aids binary detection but also helps in identifying the pixels modified by the GM. The template is learned by leveraging local and global-level features estimated by a two-branch architecture. We show that MaLP performs better than prior passive works. We also show the generalizability of MaLP by testing on 22 different GMs, providing a benchmark for future research on manipulation localization. Finally, we show that MaLP can be used as a discriminator for improving the generation quality of GMs. Our models/codes are available at www.github.com/vishal3477/pro_loc.
****
Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models
Patrick Schramowski, Manuel Brack, Björn Deiseroth, Kristian Kersting; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22522-22531
Text-conditioned image generation models have recently achieved astonishing results in image quality and text alignment and are consequently employed in a fast-growing number of applications. Since they are highly data-driven, relying on billion-sized datasets randomly scraped from the internet, they also suffer, as we demonstrate, from degenerated and biased human behavior. In turn, they may even reinforce such biases. To help combat these undesired side effects, we present safe latent diffusion (SLD). Specifically, to measure the inappropriate degeneration due to unfiltered and imbalanced training sets, we establish a novel image generation test bed - inappropriate image prompts (I2P) - containing dedicated, real-world image-to-text prompts covering concepts such as nudity and violence. As our exhaustive empirical evaluation demonstrates, the introduced SLD removes and suppresses inappropriate image parts during the diffusion process, with no additional training required and no adverse effect on overall image quality or text alignment.
****
MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation
Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10219-10228
We propose the first joint audio-video generation framework that brings engaging watching and listening experiences simultaneously, towards high-quality realistic videos. To generate joint audio-video pairs, we propose a novel Multi-Modal Diffusion model (i.e., MM-Diffusion), with two-coupled denoising autoencoders. In contrast to existing single-modal diffusion models, MM-Diffusion consists of a sequential multi-modal U-Net for a joint denoising process by design. Two subnet

s for audio and video learn to gradually generate aligned audio-video pairs from Gaussian noises. To ensure semantic consistency across modalities, we propose a novel random-shift based attention block bridging over the two subnets, which enables efficient cross-modal alignment, and thus reinforces the audio-video fidelity for each other. Extensive experiments show superior results in unconditional audio-video generation, and zero-shot conditional tasks (e.g., video-to-audio). In particular, we achieve the best FVD and FAD on Landscape and AIST++ dancing datasets. Turing tests of 10k votes further demonstrate dominant preferences for our model.
****
HexPlane: A Fast Representation for Dynamic Scenes
Ang Cao, Justin Johnson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 130-141
Modeling and re-rendering dynamic 3D scenes is a challenging task in 3D vision. Prior approaches build on NeRF and rely on implicit representations. This is slow since it requires many MLP evaluations, constraining real-world applications. We show that dynamic 3D scenes can be explicitly represented by six planes of learned features, leading to an elegant solution we call HexPlane. A HexPlane computes features for points in spacetime by fusing vectors extracted from each plane, which is highly efficient. Pairing a HexPlane with a tiny MLP to regress output colors and training via volume rendering gives impressive results for novel view synthesis on dynamic scenes, matching the image quality of prior work but reducing training time by more than 100x. Extensive ablations confirm our HexPlane design and show that it is robust to different feature fusion mechanisms, coordinate systems, and decoding mechanisms. HexPlane is a simple and effective solution for representing 4D volumes, and we hope they can broadly contribute to modeling spacetime for dynamic 3D scenes.
****
Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data
Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, Xuequan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7548-7557
Semi-supervised learning (SSL) has attracted enormous attention due to its vast potential of mitigating the dependence on large labeled datasets. The latest methods (e.g., FixMatch) use a combination of consistency regularization and pseudo-labeling to achieve remarkable successes. However, these methods all suffer from the waste of complicated examples since all pseudo-labels have to be selected by a high threshold to filter out noisy ones. Hence, the examples with ambiguous predictions will not contribute to the training phase. For better leveraging all unlabeled examples, we propose two novel techniques: Entropy Meaning Loss (EML) and Adaptive Negative Learning (ANL). EML incorporates the prediction distribution of non-target classes into the optimization objective to avoid competition with target class, and thus generating more high-confidence predictions for selecting pseudo-label. ANL introduces the additional negative pseudo-label for all unlabeled data to leverage low-confidence examples. It adaptively allocates this label by dynamically evaluating the top-k performance of the model. EML and ANL do not introduce any additional parameter and hyperparameter. We integrate these techniques with FixMatch, and develop a simple yet powerful framework called FullMatch. Extensive experiments on several common SSL benchmarks (CIFAR-10/100, SVHN, STL-10 and ImageNet) demonstrate that FullMatch exceeds FixMatch by a large margin. Integrated with FlexMatch (an advanced FixMatch-based framework), we achieve state-of-the-art performance. Source code is available at https://github.com/megvii-research/FullMatch.
****
Novel-View Acoustic Synthesis
Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6409-6419
We introduce the novel-view acoustic synthesis (NVAS) task: given the sight and sound observed at a source viewpoint, can we synthesize the sound of that scene

from an unseen target viewpoint? We propose a neural rendering approach: Visuall y-Guided Acoustic Synthesis (ViGAS) network that learns to synthesize the sound of an arbitrary point in space by analyzing the input audio-visual cues. To benc hmark this task, we collect two first-of-their-kind large-scale multi-view audio -visual datasets, one synthetic and one real. We show that our model successfull y reasons about the spatial cues and synthesizes faithful audio on both datasets . To our knowledge, this work represents the very first formulation, dataset, an d approach to solve the novel-view acoustic synthesis task, which has exciting p otential applications ranging from AR/VR to art and design. Unlocked by this wor k, we believe that the future of novel-view synthesis is in multi-modal learning from videos.
****
Robust Generalization Against Photon-Limited Corruptions via Worst-Case Sharpnes s Minimization
Zhuo Huang, Miaoxi Zhu, Xiaobo Xia, Li Shen, Jun Yu, Chen Gong, Bo Han, Bo Du, T ongliang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2023, pp. 16175-16185
Robust generalization aims to tackle the most challenging data distributions whi ch are rare in the training set and contain severe noises, i.e., photon-limited corruptions. Common solutions such as distributionally robust optimization (DRO) focus on the worst-case empirical risk to ensure low training error on the unco mmon noisy distributions. However, due to the over-parameterized model being opt imized on scarce worst-case data, DRO fails to produce a smooth loss landscape, thus struggling on generalizing well to the test set. Therefore, instead of focu sing on the worst-case risk minimization, we propose SharpDRO by penalizing the sharpness of the worst-case distribution, which measures the loss changes around the neighbor of learning parameters. Through worst-case sharpness minimization, the proposed method successfully produces a flat loss curve on the corrupted di stributions, thus achieving robust generalization. Moreover, by considering whet her the distribution annotation is available, we apply SharpDRO to two problem s ettings and design a worst-case selection process for robust generalization. The oretically, we show that SharpDRO has a great convergence guarantee. Experimenta lly, we simulate photon-limited corruptions using CIFAR10/100 and ImageNet30 dat asets and show that SharpDRO exhibits a strong generalization ability against se vere corruptions and exceeds well-known baseline methods with large performance gains.
****
Point2Pix: Photo-Realistic Point Cloud Rendering via Neural Radiance Fields
Tao Hu, Xiaogang Xu, Shu Liu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8349-8358
Synthesizing photo-realistic images from a point cloud is challenging because of the sparsity of point cloud representation. Recent Neural Radiance Fields and e xtensions are proposed to synthesize realistic images from 2D input. In this pap er, we present Point2Pix as a novel point renderer to link the 3D sparse point c louds with 2D dense image pixels. Taking advantage of the point cloud 3D prior a nd NeRF rendering pipeline, our method can synthesize high-quality images from c olored point clouds, generally for novel indoor scenes. To improve the efficienc y of ray sampling, we propose point-guided sampling, which focuses on valid samp les. Also, we present Point Encoding to build Multi-scale Radiance Fields that p rovide discriminative 3D point features. Finally, we propose Fusion Encoding to efficiently synthesize high-quality images. Extensive experiments on the ScanNet and ArkitScenes datasets demonstrate the effectiveness and generalization.
****
Superclass Learning With Representation Enhancement
Zeyu Gan, Suyun Zhao, Jinlong Kang, Liyuan Shang, Hong Chen, Cuiping Li; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24060-24069
In many real scenarios, data are often divided into a handful of artificial supe r categories in terms of expert knowledge rather than the representations of ima ges. Concretely, a superclass may contain massive and various raw categories, su

ch as refuse sorting. Due to the lack of common semantic features, the existing classification techniques are intractable to recognize superclass without raw class labels, thus they suffer severe performance damage or require huge annotation costs. To narrow this gap, this paper proposes a superclass learning framework, called SuperClass Learning with Representation Enhancement(SCLRE), to recognize super categories by leveraging enhanced representation. Specifically, by exploiting the self-attention technique across the batch, SCLRE collapses the boundaries of those raw categories and enhances the representation of each superclass. On the enhanced representation space, a superclass-aware decision boundary is then reconstructed. Theoretically, we prove that by leveraging attention techniques the generalization error of SCLRE can be bounded under superclass scenarios. Experimentally, extensive results demonstrate that SCLRE outperforms the baseline and other contrastive-based methods on CIFAR-100 datasets and four high-resolution datasets.
****
Visual Prompt Tuning for Generative Transfer Learning
Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, Lu Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19840-19851
Learning generative image models from various domains efficiently needs transferring knowledge from an image synthesis model trained on a large dataset. We present a recipe for learning vision transformers by generative knowledge transfer. We base our framework on generative vision transformers representing an image as a sequence of visual tokens with the autoregressive or non-autoregressive transformers. To adapt to a new domain, we employ prompt tuning, which prepends learnable tokens called prompts to the image token sequence and introduces a new prompt design for our task. We study on a variety of visual domains with varying amounts of training images. We show the effectiveness of knowledge transfer and a significantly better image generation quality. Code is available at https://github.com/google-research/generative_transfer.
****
NICO++: Towards Better Benchmarking for Domain Generalization
Xingxuan Zhang, Yue He, Renzhe Xu, Han Yu, Zheyan Shen, Peng Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16036-16047
Despite the remarkable performance that modern deep neural networks have achieved on independent and identically distributed (I.I.D.) data, they can crash under distribution shifts. Most current evaluation methods for domain generalization (DG) adopt the leave-one-out strategy as a compromise on the limited number of domains. We propose a large-scale benchmark with extensive labeled domains named NICO++ along with more rational evaluation methods for comprehensively evaluating DG algorithms. To evaluate DG datasets, we propose two metrics to quantify covariate shift and concept shift, respectively. Two novel generalization bounds from the perspective of data construction are proposed to prove that limited concept shift and significant covariate shift favor the evaluation capability for generalization. Through extensive experiments, NICO++ shows its superior evaluation capability compared with current DG datasets and its contribution in alleviating unfairness caused by the leak of oracle knowledge in model selection.
****
CHMATCH: Contrastive Hierarchical Matching and Robust Adaptive Threshold Boosted Semi-Supervised Learning
Jianlong Wu, Haozhe Yang, Tian Gan, Ning Ding, Feijun Jiang, Liqiang Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15762-15772
The recently proposed FixMatch and FlexMatch have achieved remarkable results in the field of semi-supervised learning. But these two methods go to two extremes as FixMatch and FlexMatch use a pre-defined constant threshold for all classes and an adaptive threshold for each category, respectively. By only investigating consistency regularization, they also suffer from unstable results and indiscriminative feature representation, especially under the situation of few labeled s

amples. In this paper, we propose a novel CHMatch method, which can learn robust adaptive thresholds for instance-level prediction matching as well as discriminative features by contrastive hierarchical matching. We first present a memory-bank based robust threshold learning strategy to select highly-confident samples. In the meantime, we make full use of the structured information in the hierarchical labels to learn an accurate affinity graph for contrastive learning. CHMatch achieves very stable and superior results on several commonly-used benchmarks. For example, CHMatch achieves 8.44% and 9.02% error rate reduction over FlexMatch on CIFAR-100 under WRN-28-2 with only 4 and 25 labeled samples per class, respectively.
****
Neural Dependencies Emerging From Learning Massive Categories
Ruili Feng, Kecheng Zheng, Kai Zhu, Yujun Shen, Jian Zhao, Yukun Huang, Deli Zhao, Jingren Zhou, Michael Jordan, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11711-11720
This work presents two astonishing findings on neural networks learned for large-scale image classification. 1) Given a well-trained model, the logits predicted for some category can be directly obtained by linearly combining the predictions of a few other categories, which we call neural dependency. 2) Neural dependencies exist not only within a single model, but even between two independently learned models, regardless of their architectures. Towards a theoretical analysis of such phenomena, we demonstrate that identifying neural dependencies is equivalent to solving the Covariance Lasso (CovLasso) regression problem proposed in this paper. Through investigating the properties of the problem solution, we confirm that neural dependency is guaranteed by a redundant logit covariance matrix, which condition is easily met given massive categories, and that neural dependency is sparse, which implies one category relates to only a few others. We further empirically show the potential of neural dependencies in understanding internal data correlations, generalizing models to unseen categories, and improving model robustness with a dependency-derived regularize. Code to exactly reproduce the results in this work will be released publicly.
****
ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects
Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, Samuele Salti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20762-20772
In this paper, we focus on the problem of rendering novel views from a Neural Radiance Field (NeRF) under unobserved light conditions. To this end, we introduce a novel dataset, dubbed ReNe (Relighting NeRF), framing real world objects under one-light-at-time (OLAT) conditions, annotated with accurate ground-truth camera and light poses. Our acquisition pipeline leverages two robotic arms holding, respectively, a camera and an omni-directional point-wise light source. We release a total of 20 scenes depicting a variety of objects with complex geometry and challenging materials. Each scene includes 2000 images, acquired from 50 different points of views under 40 different OLAT conditions. By leveraging the dataset, we perform an ablation study on the relighting capability of variants of the vanilla NeRF architecture and identify a lightweight architecture that can render novel views of an object under novel light conditions, which we use to establish a non-trivial baseline for the dataset. Dataset and benchmark are available at https://eyecan-ai.github.io/rene.
****
ARCTIC: A Dataset for Dexterous Bimanual Hand-Object Manipulation
Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12943-12954
Humans intuitively understand that inanimate objects do not move by themselves, but that state changes are typically caused by human manipulation (e.g., the opening of a book). This is not yet the case for machines. In part this is because there exist no datasets with ground-truth 3D annotations for the study of physic

ally consistent and synchronised motion of hands and articulated objects. To this end, we introduce ARCTIC -- a dataset of two hands that dexterously manipulate objects, containing 2.1M video frames paired with accurate 3D hand and object meshes and detailed, dynamic contact information. It contains bi-manual articulation of objects such as scissors or laptops, where hand poses and object states evolve jointly in time. We propose two novel articulated hand-object interaction tasks: (1) Consistent motion reconstruction: Given a monocular video, the goal is to reconstruct two hands and articulated objects in 3D, so that their motions are spatio-temporally consistent. (2) Interaction field estimation: Dense relative hand-object distances must be estimated from images. We introduce two baselines ArcticNet and InterField, respectively and evaluate them qualitatively and quantitatively on ARCTIC. Our code and data are available at https://arctic.is.tue.mpg.de.
****
Constrained Evolutionary Diffusion Filter for Monocular Endoscope Tracking
Xiongbiao Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4747-4756
Stochastic filtering is widely used to deal with nonlinear optimization problems such as 3-D and visual tracking in various computer vision and augmented reality applications. Many current methods suffer from an imbalance between exploration and exploitation due to their particle degeneracy and impoverishment, resulting in local optimums. To address this imbalance, this work proposes a new constrained evolutionary diffusion filter for nonlinear optimization. Specifically, this filter develops spatial state constraints and adaptive history-recall differential evolution embedded evolutionary stochastic diffusion instead of sequential resampling to resolve the degeneracy and impoverishment problem. With application to monocular endoscope 3-D tracking, the experimental results show that the proposed filtering significantly improves the balance between exploration and exploitation and certainly works better than recent 3-D tracking methods. Particularly, the surgical tracking error was reduced from 4.03 mm to 2.59 mm.
****
MAGVIT: Masked Generative Video Transformer
Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G. Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, Lu Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10459-10469
We introduce the MAsked Generative VIdeo Transformer, MAGVIT, to tackle various video synthesis tasks with a single model. We introduce a 3D tokenizer to quantize a video into spatial-temporal visual tokens and propose an embedding method for masked video token modeling to facilitate multi-task learning. We conduct extensive experiments to demonstrate the quality, efficiency, and flexibility of MAGVIT. Our experiments show that (i) MAGVIT performs favorably against state-of-the-art approaches and establishes the best-published FVD on three video generation benchmarks, including the challenging Kinetics-600. (ii) MAGVIT outperforms existing methods in inference time by two orders of magnitude against diffusion models and by 60x against autoregressive models. (iii) A single MAGVIT model supports ten diverse generation tasks and generalizes across videos from different visual domains. The source code and trained models will be released to the public at https://magvit.cs.cmu.edu.
****
Content-Aware Token Sharing for Efficient Semantic Segmentation With Vision Transformers
Chenyang Lu, Daan de Geus, Gijs Dubbelman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23631-23640
This paper introduces Content-aware Token Sharing (CTS), a token reduction approach that improves the computational efficiency of semantic segmentation networks that use Vision Transformers (ViTs). Existing works have proposed token reduction approaches to improve the efficiency of ViT-based image classification networks, but these methods are not directly applicable to semantic segmentation, which we address in this work. We observe that, for semantic segmentation, multiple

image patches can share a token if they contain the same semantic class, as they contain redundant information. Our approach leverages this by employing an efficient, class-agnostic policy network that predicts if image patches contain the same semantic class, and lets them share a token if they do. With experiments, we explore the critical design choices of CTS and show its effectiveness on the ADE20K, Pascal Context and Cityscapes datasets, various ViT backbones, and different segmentation decoders. With Content-aware Token Sharing, we are able to reduce the number of processed tokens by up to 44%, without diminishing the segmentation quality.
****
Toward Accurate Post-Training Quantization for Image Super Resolution
Zhijun Tu, Jie Hu, Hanting Chen, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5856-5865
Model quantization is a crucial step for deploying super resolution (SR) networks on mobile devices. However, existing works focus on quantization-aware training, which requires complete dataset and expensive computational overhead. In this paper, we study post-training quantization(PTQ) for image super resolution using only a few unlabeled calibration images. As the SR model aims to maintain the texture and color information of input images, the distribution of activations are long-tailed, asymmetric and highly dynamic compared with classification models. To this end, we introduce the density-based dual clipping to cut off the outliers based on analyzing the asymmetric bounds of activations. Moreover, we present a novel pixel aware calibration method with the supervision of the full-precision model to accommodate the highly dynamic range of different samples. Extensive experiments demonstrate that the proposed method significantly outperforms existing PTQ algorithms on various models and datasets. For instance, we get a 2.091 dB increase on Urban100 benchmark when quantizing EDSRx4 to 4-bit with 100 unlabeled images. Our code is available at both https://github.com/huawei-noah/Efficient-Computing/tree/master/Quantization/PTQ4SR and https://gitee.com/mindspore/models/tree/master/research/cv/PTQ4SR.
****
Hidden Gems: 4D Radar Scene Flow Learning Using Cross-Modal Supervision
Fangqiang Ding, Andras Palffy, Dariu M. Gavrila, Chris Xiaoxuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9340-9349
This work proposes a novel approach to 4D radar-based scene flow estimation via cross-modal learning. Our approach is motivated by the co-located sensing redundancy in modern autonomous vehicles. Such redundancy implicitly provides various forms of supervision cues to the radar scene flow estimation. Specifically, we introduce a multi-task model architecture for the identified cross-modal learning problem and propose loss functions to opportunistically engage scene flow estimation using multiple cross-modal constraints for effective model training. Extensive experiments show the state-of-the-art performance of our method and demonstrate the effectiveness of cross-modal supervised learning to infer more accurate 4D radar scene flow. We also show its usefulness to two subtasks - motion segmentation and ego-motion estimation. Our source code will be available on https://github.com/Toytiny/CMFlow.
****
OmniMAE: Single Model Masked Pretraining on Images and Videos
Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10406-10417
Transformer-based architectures have become competitive across a variety of visual domains, most notably images and videos. While prior work studies these modalities in isolation, having a common architecture suggests that one can train a single unified model for multiple visual modalities. Prior attempts at unified modeling typically use architectures tailored for vision tasks, or obtain worse performance compared to single modality models. In this work, we show that masked autoencoding can be used to train a simple Vision Transformer on images and videos, without requiring any labeled data. This single model learns visual represen

tations that are comparable to or better than single-modality representations on
 both image and video benchmarks, while using a much simpler architecture. Furth
ermore, this model can be learned by dropping 90% of the image and 95% of the vi
deo patches, enabling extremely fast training of huge model architectures. In pa
rticular, we show that our single ViT-Huge model can be finetuned to achieve 86.
6% on ImageNet and 75.5% on the challenging Something Something-v2 video benchma
rk, setting a new state-of-the-art.
****
Omnimatte3D: Associating Objects and Their Effects in Unconstrained Monocular Vi
deo
Mohammed Suhail, Erika Lu, Zhengqi Li, Noah Snavely, Leonid Sigal, Forrester Col
e; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2023, pp. 630-639
We propose a method to decompose a video into a background and a set of foregrou
nd layers, where the background captures stationary elements while the foregroun
d layers capture moving objects along with their associated effects (e.g. shadow
s and reflections). Our approach is designed for unconstrained monocular videos,
 with arbitrary camera and object motion. Prior work that tackles this problem a
ssumes that the video can be mapped onto a fixed 2D canvas, severely limiting th
e possible space of camera motion. Instead, our method applies recent progress i
n monocular camera pose and depth estimation to create a full, RGBD video layer
for the background, along with a video layer for each foreground object. To solv
e the underconstrained decomposition problem, we propose a new loss formulation
based on multi-view consistency. We test our method on challenging videos with c
omplex camera motion and show significant qualitative improvement over current a
pproaches.
****
Real-Time Neural Light Field on Mobile Devices
Junli Cao, Huan Wang, Pavlo Chemerys, Vladislav Shakhrai, Ju Hu, Yun Fu, Denys M
akoviichuk, Sergey Tulyakov, Jian Ren; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8328-8337
Recent efforts in Neural Rendering Fields (NeRF) have shown impressive results o
n novel view synthesis by utilizing implicit neural representation to represent
3D scenes. Due to the process of volumetric rendering, the inference speed for N
eRF is extremely slow, limiting the application scenarios of utilizing NeRF on r
esource-constrained hardware, such as mobile devices. Many works have been condu
cted to reduce the latency of running NeRF models. However, most of them still r
equire high-end GPU for acceleration or extra storage memory, which is all unava
ilable on mobile devices. Another emerging direction utilizes the neural light f
ield (NeLF) for speedup, as only one forward pass is performed on a ray to predi
ct the pixel color. Nevertheless, to reach a similar rendering quality as NeRF,
the network in NeLF is designed with intensive computation, which is not mobile-
friendly. In this work, we propose an efficient network that runs in real-time o
n mobile devices for neural rendering. We follow the setting of NeLF to train ou
r network. Unlike existing works, we introduce a novel network architecture that
 runs efficiently on mobile devices with low latency and small size, i.e., savin
g 15x   24x storage compared with MobileNeRF. Our model achieves high-resolution
 generation while maintaining real-time inference for both synthetic and real-wo
rld scenes on mobile devices, e.g., 18.04ms (iPhone 13) for rendering one 1008x7
56 image of real 3D scenes. Additionally, we achieve similar image quality as Ne
RF and better quality than MobileNeRF (PSNR 26.15 vs. 25.91 on the real-world fo
rward-facing dataset).
****
Incrementer: Transformer for Class-Incremental Semantic Segmentation With Knowle
dge Distillation Focusing on Old Class
Chao Shang, Hongliang Li, Fanman Meng, Qingbo Wu, Heqian Qiu, Lanxiao Wang; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2023, pp. 7214-7224
Class-incremental semantic segmentation aims to incrementally learn new classes
while maintaining the capability to segment old ones, and suffers catastrophic f

orgetting since the old-class labels are unavailable. Most existing methods are based on convolutional networks and prevent forgetting through knowledge distillation, which (1) need to add additional convolutional layers to predict new classes, and (2) ignore to distinguish different regions corresponding to old and new classes during knowledge distillation and roughly distill all the features, thus limiting the learning of new classes. Based on the above observations, we propose a new transformer framework for class-incremental semantic segmentation, dubbed Incrementer, which only needs to add new class tokens to the transformer decoder for new-class learning. Based on the Incrementer, we propose a new knowledge distillation scheme that focuses on the distillation in the old-class regions, which reduces the constraints of the old model on the new-class learning, thus improving the plasticity. Moreover, we propose a class deconfusion strategy to alleviate the overfitting to new classes and the confusion of similar classes. Our method is simple and effective, and extensive experiments show that our method outperforms the SOTAs by a large margin (5 15 absolute points boosts on both Pascal VOC and ADE20k). We hope that our Incrementer can serve as a new strong pipeline for class-incremental semantic segmentation.
****
End-to-End Video Matting With Trimap Propagation
Wei-Lun Huang, Ming-Sui Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14337-14347
The research of video matting mainly focuses on temporal coherence and has gained significant improvement via neural networks. However, matting usually relies on user-annotated trimaps to estimate alpha values, which is a labor-intensive issue. Although recent studies exploit video object segmentation methods to propagate the given trimaps, they suffer inconsistent results. Here we present a more robust and faster end-to-end video matting model equipped with trimap propagation called FTP-VM (Fast Trimap Propagation - Video Matting). The FTP-VM combines trimap propagation and video matting in one model, where the additional backbone in memory matching is replaced with the proposed lightweight trimap fusion module. The segmentation consistency loss is adopted from automotive segmentation to fit trimap segmentation with the collaboration of RNN (Recurrent Neural Network) to improve the temporal coherence. The experimental results demonstrate that the FTP-VM performs competitively both in composited and real videos only with few given trimaps. The efficiency is eight times higher than the state-of-the-art methods, which confirms its robustness and applicability in real-time scenarios. The code is available at https://github.com/csvt32745/FTP-VM
****
DropMAE: Masked Autoencoders With Spatial-Attention Dropout for Tracking Tasks
Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14561-14571
In this paper, we study masked autoencoder (MAE) pretraining on videos for matching-based downstream tasks, including visual object tracking (VOT) and video object segmentation (VOS). A simple extension of MAE is to randomly mask out frame patches in videos and reconstruct the frame pixels. However, we find that this simple baseline heavily relies on spatial cues while ignoring temporal relations for frame reconstruction, thus leading to sub-optimal temporal matching representations for VOT and VOS. To alleviate this problem, we propose DropMAE, which adaptively performs spatial-attention dropout in the frame reconstruction to facilitate temporal correspondence learning in videos. We show that our DropMAE is a strong and efficient temporal matching learner, which achieves better finetuning results on matching-based tasks than the ImageNetbased MAE with 2x faster pre-training speed. Moreover, we also find that motion diversity in pre-training videos is more important than scene diversity for improving the performance on VOT and VOS. Our pre-trained DropMAE model can be directly loaded in existing ViT-based trackers for fine-tuning without further modifications. Notably, DropMAE sets new state-of-the-art performance on 8 out of 9 highly competitive video tracking and segmentation datasets. Our code and pre-trained models are available at https://github.com/jimmy-dq/DropMAE.git.

****
Are Binary Annotations Sufficient? Video Moment Retrieval via Hierarchical Uncertainty-Based Active Learning
Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, Tat-seng Chua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23013-23022
Recent research on video moment retrieval has mostly focused on enhancing the performance of accuracy, efficiency, and robustness, all of which largely rely on the abundance of high-quality annotations. While the precise frame-level annotations are time-consuming and cost-expensive, few attentions have been paid to the labeling process. In this work, we explore a new interactive manner to stimulate the process of human-in-the-loop annotation in video moment retrieval task. The key challenge is to select "ambiguous" frames and videos for binary annotations to facilitate the network training. To be specific, we propose a new hierarchical uncertainty-based modeling that explicitly considers modeling the uncertainty of each frame within the entire video sequence corresponding to the query description, and selecting the frame with the highest uncertainty. Only selected frame will be annotated by the human experts, which can largely reduce the workload. After obtaining a small number of labels provided by the expert, we show that it is sufficient to learn a competitive video moment retrieval model in such a harsh environment. Moreover, we treat the uncertainty score of frames in a video as a whole, and estimate the difficulty of each video, which can further relieve the burden of video selection. In general, our active learning strategy for video moment retrieval works not only at the frame level but also at the sequence level. Experiments on two public datasets validate the effectiveness of our proposed method.
****
High-Fidelity Clothed Avatar Reconstruction From a Single Image
Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, Zhen Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8662-8672
This paper presents a framework for efficient 3D clothed avatar reconstruction. By combining the advantages of the high accuracy of optimization-based methods and the efficiency of learning-based methods, we propose a coarse-to-fine way to realize a high-fidelity clothed avatar reconstruction (CAR) from a single image. At the first stage, we use an implicit model to learn the general shape in the canonical space of a person in a learning-based way, and at the second stage, we refine the surface detail by estimating the non-rigid deformation in the posed space in an optimization way. A hyper-network is utilized to generate a good initialization so that the convergence of the optimization process is greatly accelerated. Extensive experiments on various datasets show that the proposed CAR successfully produces high-fidelity avatars for arbitrarily clothed humans in real scenes. The codes will be released.
****
Zero-Shot Object Counting
Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, Dimitris Samaras; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15548-15557
Class-agnostic object counting aims to count object instances of an arbitrary class at test time. It is challenging but also enables many potential applications. Current methods require human-annotated exemplars as inputs which are often unavailable for novel categories, especially for autonomous systems. Thus, we propose zero-shot object counting (ZSC), a new setting where only the class name is available during test time. Such a counting system does not require human annotators in the loop and can operate automatically. Starting from a class name, we propose a method that can accurately identify the optimal patches which can then be used as counting exemplars. Specifically, we first construct a class prototype to select the patches that are likely to contain the objects of interest, namely class-relevant patches. Furthermore, we introduce a model that can quantitatively measure how suitable an arbitrary patch is as a counting exemplar. By apply

ing this model to all the candidate patches, we can select the most suitable patches as exemplars for counting. Experimental results on a recent class-agnostic counting dataset, FSC-147, validate the effectiveness of our method.
****
Patch-Mix Transformer for Unsupervised Domain Adaptation: A Game Perspective
Jinjing Zhu, Haotian Bai, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3561-3571
Endeavors have been recently made to leverage the vision transformer (ViT) for the challenging unsupervised domain adaptation (UDA) task. They typically adopt the cross-attention in ViT for direct domain alignment. However, as the performance of cross-attention highly relies on the quality of pseudo labels for targeted samples, it becomes less effective when the domain gap becomes large. We solve this problem from a game theory's perspective with the proposed model dubbed as PMTrans, which bridges source and target domains with an intermediate domain. Specifically, we propose a novel ViT-based module called PatchMix that effectively builds up the intermediate domain, i.e., probability distribution, by learning to sample patches from both domains based on the game-theoretical models. This way, it learns to mix the patches from the source and target domains to maximize the cross entropy (CE), while exploiting two semi-supervised mixup losses in the feature and label spaces to minimize it. As such, we interpret the process of UDA as a min-max CE game with three players, including the feature extractor, classifier, and PatchMix, to find the Nash Equilibria. Moreover, we leverage attention maps from ViT to re-weight the label of each patch by its importance, making it possible to obtain more domain-discriminative feature representations. We conduct extensive experiments on four benchmark datasets, and the results show that PMTrans significantly surpasses the ViT-based and CNN-based SoTA methods by +3.6% on Office-Home, +1.4% on Office-31, and +17.7% on DomainNet, respectively. https://vlis2022.github.io/cvpr23/PMTrans
****
Implicit Diffusion Models for Continuous Super-Resolution
Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, Baochang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10021-10030
Image super-resolution (SR) has attracted increasing attention due to its wide applications. However, current SR methods generally suffer from over-smoothing and artifacts, and most work only with fixed magnifications. This paper introduces an Implicit Diffusion Model (IDM) for high-fidelity continuous image super-resolution. IDM integrates an implicit neural representation and a denoising diffusion model in a unified end-to-end framework, where the implicit neural representation is adopted in the decoding process to learn continuous-resolution representation. Furthermore, we design a scale-controllable conditioning mechanism that consists of a low-resolution (LR) conditioning network and a scaling factor. The scaling factor regulates the resolution and accordingly modulates the proportion of the LR information and generated features in the final output, which enables the model to accommodate the continuous-resolution requirement. Extensive experiments validate the effectiveness of our IDM and demonstrate its superior performance over prior arts.
****
VGFlow: Visibility Guided Flow Network for Human Reposing
Rishabh Jain, Krishna Kumar Singh, Mayur Hemani, Jingwan Lu, Mausoom Sarkar, Duygu Ceylan, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21088-21097
The task of human reposing involves generating a realistic image of a model standing in an arbitrary conceivable pose. There are multiple difficulties in generating perceptually accurate images and existing methods suffers from limitations in preserving texture, maintaining pattern coherence, respecting cloth boundaries, handling occlusions, manipulating skin generation etc. These difficulties are further exacerbated by the fact that the possible space of pose orientation for humans is large and variable, the nature of clothing items are highly non-rigid and the diversity in body shape differ largely among the population. To allevia

te these difficulties and synthesize perceptually accurate images, we propose VGFlow, a model which uses a visibility guided flow module to disentangle the flow into visible and invisible parts of the target for simultaneous texture preservation and style manipulation. Furthermore, to tackle distinct body shapes and avoid network artifacts, we also incorporate an a self-supervised patch-wise "realness" loss to further improve the output. VGFlow achieves state-of-the-art results as observed qualitatively and quantitatively on different image quality metrics(SSIM, LPIPS, FID).
****
Phase-Shifting Coder: Predicting Accurate Orientation in Oriented Object Detection
Yi Yu, Feipeng Da; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13354-13363
With the vigorous development of computer vision, oriented object detection has gradually been featured. In this paper, a novel differentiable angle coder named phase-shifting coder (PSC) is proposed to accurately predict the orientation of objects, along with a dual-frequency version (PSCD). By mapping the rotational periodicity of different cycles into the phase of different frequencies, we provide a unified framework for various periodic fuzzy problems caused by rotational symmetry in oriented object detection. Upon such a framework, common problems in oriented object detection such as boundary discontinuity and square-like problems are elegantly solved in a unified form. Visual analysis and experiments on three datasets prove the effectiveness and the potentiality of our approach. When facing scenarios requiring high-quality bounding boxes, the proposed methods are expected to give a competitive performance. The codes are publicly available at https://github.com/open-mmlab/mmrotate.
****
Improving Selective Visual Question Answering by Learning From Your Peers
Corentin Dancette, Spencer Whitehead, Rishabh Maheshwary, Ramakrishna Vedantam, Stefan Scherer, Xinlei Chen, Matthieu Cord, Marcus Rohrbach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24049-24059
Despite advances in Visual Question Answering (VQA), the ability of models to assess their own correctness remains underexplored. Recent work has shown that VQA models, out-of-the-box, can have difficulties abstaining from answering when they are wrong. The option to abstain, also called Selective Prediction, is highly relevant when deploying systems to users who must trust the system's output (e.g., VQA assistants for users with visual impairments). For such scenarios, abstention can be especially important as users may provide out-of-distribution (OOD) or adversarial inputs that make incorrect answers more likely. In this work, we explore Selective VQA in both in-distribution (ID) and OOD scenarios, where models are presented with mixtures of ID and OOD data. The goal is to maximize the number of questions answered while minimizing the risk of error on those questions. We propose a simple yet effective Learning from Your Peers (LYP) approach for training multimodal selection functions for making abstention decisions. Our approach uses predictions from models trained on distinct subsets of the training data as targets for optimizing a Selective VQA model. It does not require additional manual labels or held-out data and provides a signal for identifying examples that are easy/difficult to generalize to. In our extensive evaluations, we show this benefits a number of models across different architectures and scales. Overall, for ID, we reach 32.92% in the selective prediction metric coverage at 1% risk of error (C@1%) which doubles the previous best coverage of 15.79% on this task. For mixed ID/OOD, using models' softmax confidences for abstention decisions performs very poorly, answering <5% of questions at 1% risk of error even when faced with only 10% OOD examples, but a learned selection function with LYP can increase that to 25.38% C@1%.
****
CAMS: CAnonicalized Manipulation Spaces for Category-Level Functional Hand-Object Manipulation Synthesis
Juntian Zheng, Qingyuan Zheng, Lixing Fang, Yun Liu, Li Yi; Proceedings of the I

EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 585-594
In this work, we focus on a novel task of category-level functional hand-object manipulation synthesis covering both rigid and articulated object categories. Given an object geometry, an initial human hand pose as well as a sparse control sequence of object poses, our goal is to generate a physically reasonable hand-object manipulation sequence that performs like human beings. To address such a challenge, we first design CAnonicalized Manipulation Spaces (CAMS), a two-level space hierarchy that canonicalizes the hand poses in an object-centric and contact-centric view. Benefiting from the representation capability of CAMS, we then present a two-stage framework for synthesizing human-like manipulation animations. Our framework achieves state-of-the-art performance for both rigid and articulated categories with impressive visual effects. Codes and video results can be found at our project homepage: https://cams-hoi.github.io/.
****

Neural Lens Modeling
Wenqi Xian, Aljaž Boži■, Noah Snavely, Christoph Lassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8435-8445
Recent methods for 3D reconstruction and rendering increasingly benefit from end-to-end optimization of the entire image formation process. However, this approach is currently limited: effects of the optical hardware stack and in particular lenses are hard to model in a unified way. This limits the quality that can be achieved for camera calibration and the fidelity of the results of 3D reconstruction. In this paper, we propose NeuroLens, a neural lens model for distortion and vignetting that can be used for point projection and ray casting and can be optimized through both operations. This means that it can (optionally) be used to perform pre-capture calibration using classical calibration targets, and can later be used to perform calibration or refinement during 3D reconstruction, e.g., while optimizing a radiance field. To evaluate the performance of our proposed model, we create a comprehensive dataset assembled from the Lensfun database with a multitude of lenses. Using this and other real-world datasets, we show that the quality of our proposed lens model outperforms standard packages as well as recent approaches while being much easier to use and extend. The model generalizes across many lens types and is trivial to integrate into existing 3D reconstruction and rendering systems. Visit our project website at: https://neural-lens.github.io.
****

CoralStyleCLIP: Co-Optimized Region and Layer Selection for Image Editing
Ambareesh Revanur, Debraj Basu, Shradha Agrawal, Dhwanit Agarwal, Deepak Pai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12695-12704
Edit fidelity is a significant issue in open-world controllable generative image editing. Recently, CLIP-based approaches have traded off simplicity to alleviate these problems by introducing spatial attention in a handpicked layer of a StyleGAN. In this paper, we propose CoralStyleCLIP, which incorporates a multi-layer attention-guided blending strategy in the feature space of StyleGAN2 for obtaining high-fidelity edits. We propose multiple forms of our co-optimized region and layer selection strategy to demonstrate the variation of time complexity with the quality of edits over different architectural intricacies while preserving simplicity. We conduct extensive experimental analysis and benchmark our method against state-of-the-art CLIP-based methods. Our findings suggest that CoralStyleCLIP results in high-quality edits while preserving the ease of use.
****

GLeaD: Improving GANs With a Generator-Leading Task
Qingyan Bai, Ceyuan Yang, Yinghao Xu, Xihui Liu, Yujiu Yang, Yujun Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12094-12104
Generative adversarial network (GAN) is formulated as a two-player game between a generator (G) and a discriminator (D), where D is asked to differentiate wheth

er an image comes from real data or is produced by G. Under such a formulation, D plays as the rule maker and hence tends to dominate the competition. Towards a fairer game in GANs, we propose a new paradigm for adversarial training, which makes G assign a task to D as well. Specifically, given an image, we expect D to extract representative features that can be adequately decoded by G to reconstruct the input. That way, instead of learning freely, D is urged to align with the view of G for domain classification. Experimental results on various datasets demonstrate the substantial superiority of our approach over the baselines. For instance, we improve the FID of StyleGAN2 from 4.30 to 2.55 on LSUN Bedroom and from 4.04 to 2.82 on LSUN Church. We believe that the pioneering attempt present in this work could inspire the community with better designed generator-leading tasks for GAN improvement. Project page is at https://ezioby.github.io/glead/.
****
GALIP: Generative Adversarial CLIPs for Text-to-Image Synthesis
Ming Tao, Bing-Kun Bao, Hao Tang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14214-14223
Synthesizing high-fidelity complex images from text is challenging. Based on large pretraining, the autoregressive and diffusion models can synthesize photo-realistic images. Although these large models have shown notable progress, there remain three flaws. 1) These models require tremendous training data and parameters to achieve good performance. 2) The multi-step generation design slows the image synthesis process heavily. 3) The synthesized visual features are challenging to control and require delicately designed prompts. To enable high-quality, efficient, fast, and controllable text-to-image synthesis, we propose Generative Adversarial CLIPs, namely GALIP. GALIP leverages the powerful pretrained CLIP model both in the discriminator and generator. Specifically, we propose a CLIP-based discriminator. The complex scene understanding ability of CLIP enables the discriminator to accurately assess the image quality. Furthermore, we propose a CLIP-empowered generator that induces the visual concepts from CLIP through bridge features and prompts. The CLIP-integrated generator and discriminator boost training efficiency, and as a result, our model only requires about 3% training data and 6% learnable parameters, achieving comparable results to large pretrained autoregressive and diffusion models. Moreover, our model achieves 120 times faster synthesis speed and inherits the smooth latent space from GAN. The extensive experimental results demonstrate the excellent performance of our GALIP. Code is available at https://github.com/tobran/GALIP.
****
Look, Radiate, and Learn: Self-Supervised Localisation via Radio-Visual Correspondence
Mohammed Alloulah, Maximilian Arnold; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17430-17440
Next generation cellular networks will implement radio sensing functions alongside customary communications, thereby enabling unprecedented worldwide sensing coverage outdoors. Deep learning has revolutionised computer vision but has had limited application to radio perception tasks, in part due to lack of systematic datasets and benchmarks dedicated to the study of the performance and promise of radio sensing. To address this gap, we present MaxRay: a synthetic radio-visual dataset and benchmark that facilitate precise target localisation in radio. We further propose to learn to localise targets in radio without supervision by extracting self-coordinates from radio-visual correspondence. We use such self-supervised coordinates to train a radio localiser network. We characterise our performance against a number of state-of-the-art baselines. Our results indicate that accurate radio target localisation can be automatically learned from paired radio-visual data without labels, which is important for empirical data. This opens the door for vast data scalability and may prove key to realising the promise of robust radio sensing atop a unified communication-perception cellular infrastructure. Dataset will be hosted on IEEE DataPort.
****
Multiplicative Fourier Level of Detail
Yishun Dou, Zhong Zheng, Qiaoqiao Jin, Bingbing Ni; Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1808-181
7
We develop a simple yet surprisingly effective implicit representing scheme call
ed Multiplicative Fourier Level of Detail (MFLOD) motivated by the recent succes
s of multiplicative filter network. Built on multi-resolution feature grid/volum
e (e.g., the sparse voxel octree), each level's feature is first modulated by a
sinusoidal function and then element-wisely multiplied by a linear transformatio
n of previous layer's representation in a layer-to-layer recursive manner, yield
ing the scale-aggregated encodings for a subsequent simple linear forward to get
 final output. In contrast to previous hybrid representations relying on interle
aved multilevel fusion and nonlinear activation-based decoding, MFLOD could be e
legantly characterized as a linear combination of sine basis functions with vary
ing amplitude, frequency, and phase upon the learned multilevel features, thus o
ffering great feasibility in Fourier analysis. Comprehensive experimental result
s on implicit neural representation learning tasks including image fitting, 3D s
hape representation, and neural radiance fields well demonstrate the superior qu
ality and generalizability achieved by the proposed MFLOD scheme.
****
Indiscernible Object Counting in Underwater Scenes
Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, Lu
c Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2023, pp. 13791-13801
Recently, indiscernible scene understanding has attracted a lot of attention in
the vision community. We further advance the frontier of this field by systemati
cally studying a new challenge named indiscernible object counting (IOC), the go
al of which is to count objects that are blended with respect to their surroundi
ngs. Due to a lack of appropriate IOC datasets, we present a large-scale dataset
 IOCfish5K which contains a total of 5,637 high-resolution images and 659,024 an
notated center points. Our dataset consists of a large number of indiscernible o
bjects (mainly fish) in underwater scenes, making the annotation process all the
 more challenging. IOCfish5K is superior to existing datasets with indiscernible
 scenes because of its larger scale, higher image resolutions, more annotations,
 and denser scenes. All these aspects make it the most challenging dataset for I
OC so far, supporting progress in this area. For benchmarking purposes, we selec
t 14 mainstream methods for object counting and carefully evaluate them on IOCfi
sh5K. Furthermore, we propose IOCFormer, a new strong baseline that combines den
sity and regression branches in a unified framework and can effectively tackle o
bject counting under concealed scenes. Experiments show that IOCFormer achieves
state-of-the-art scores on IOCfish5K.
****
Shape-Erased Feature Learning for Visible-Infrared Person Re-Identification
Jiawei Feng, Ancong Wu, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22752-22761
Due to the modality gap between visible and infrared images with high visual amb
iguity, learning diverse modality-shared semantic concepts for visible-infrared
person re-identification (VI-ReID) remains a challenging problem. Body shape is
one of the significant modality-shared cues for VI-ReID. To dig more diverse mod
ality-shared cues, we expect that erasing body-shape-related semantic concepts i
n the learned features can force the ReID model to extract more and other modali
ty-shared features for identification. To this end, we propose shape-erased feat
ure learning paradigm that decorrelates modality-shared features in two orthogon
al subspaces. Jointly learning shape-related feature in one subspace and shape-e
rased features in the orthogonal complement achieves a conditional mutual inform
ation maximization between shape-erased feature and identity discarding body sha
pe information, thus enhancing the diversity of the learned representation expli
citly. Extensive experiments on SYSU-MM01, RegDB, and HITSZ-VCM datasets demonst
rate the effectiveness of our method.
****
Relational Context Learning for Human-Object Interaction Detection
Sanghyun Kim, Deunsol Jung, Minsu Cho; Proceedings of the IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2925-2934
Recent state-of-the-art methods for HOI detection typically build on transformer architectures with two decoder branches, one for human-object pair detection and the other for interaction classification. Such disentangled transformers, however, may suffer from insufficient context exchange between the branches and lead to a lack of context information for relational reasoning, which is critical in discovering HOI instances. In this work, we propose the multiplex relation network (MUREN) that performs rich context exchange between three decoder branches using unary, pairwise, and ternary relations of human, object, and interaction tokens. The proposed method learns comprehensive relational contexts for discovering HOI instances, achieving state-of-the-art performance on two standard benchmarks for HOI detection, HICO-DET and V-COCO.
****
Low-Light Image Enhancement via Structure Modeling and Guidance
Xiaogang Xu, Ruixing Wang, Jiangbo Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9893-9903
This paper proposes a new framework for low-light image enhancement by simultaneously conducting the appearance as well as structure modeling. It employs the structural feature to guide the appearance enhancement, leading to sharp and realistic results. The structure modeling in our framework is implemented as the edge detection in low-light images. It is achieved with a modified generative model via designing a structure-aware feature extractor and generator. The detected edge maps can accurately emphasize the essential structural information, and the edge prediction is robust towards the noises in dark areas. Moreover, to improve the appearance modeling, which is implemented with a simple U-Net, a novel structure-guided enhancement module is proposed with structure-guided feature synthesis layers. The appearance modeling, edge detector, and enhancement module can be trained end-to-end. The experiments are conducted on representative datasets (sRGB and RAW domains), showing that our model consistently achieves SOTA performance on all datasets with the same architecture.
****
On Calibrating Semantic Segmentation Models: Analyses and an Algorithm
Dongdong Wang, Boqing Gong, Liqiang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23652-23662
We study the problem of semantic segmentation calibration. Lots of solutions have been proposed to approach model miscalibration of confidence in image classification. However, to date, confidence calibration research on semantic segmentation is still limited. We provide a systematic study on the calibration of semantic segmentation models and propose a simple yet effective approach. First, we find that model capacity, crop size, multi-scale testing, and prediction correctness have impact on calibration. Among them, prediction correctness, especially misprediction, is more important to miscalibration due to over-confidence. Next, we propose a simple, unifying, and effective approach, namely selective scaling, by separating correct/incorrect prediction for scaling and more focusing on misprediction logit smoothing. Then, we study popular existing calibration methods and compare them with selective scaling on semantic segmentation calibration. We conduct extensive experiments with a variety of benchmarks on both in-domain and domain-shift calibration and show that selective scaling consistently outperforms other methods.
****
Visual Atoms: Pre-Training Vision Transformers With Sinusoidal Waves
Sora Takashima, Ryo Hayamizu, Nakamasa Inoue, Hirokatsu Kataoka, Rio Yokota; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18579-18588
Formula-driven supervised learning (FDSL) has been shown to be an effective method for pre-training vision transformers, where ExFractalDB-21k was shown to exceed the pre-training effect of ImageNet-21k. These studies also indicate that contours mattered more than textures when pre-training vision transformers. However, the lack of a systematic investigation as to why these contour-oriented synthetic datasets can achieve the same accuracy as real datasets leaves much room for

skepticism. In the present work, we develop a novel methodology based on circular harmonics for systematically investigating the design space of contour-oriented synthetic datasets. This allows us to efficiently search the optimal range of FDSL parameters and maximize the variety of synthetic images in the dataset, which we found to be a critical factor. When the resulting new dataset VisualAtom-21k is used for pre-training ViT-Base, the top-1 accuracy reached 83.7% when fine-tuning on ImageNet-1k. This is only 0.5% difference from the top-1 accuracy (84.2%) achieved by the JFT-300M pre-training, even though the scale of images is 1/14. Unlike JFT-300M which is a static dataset, the quality of synthetic datasets will continue to improve, and the current work is a testament to this possibility. FDSL is also free of the common issues associated with real images, e.g. privacy/copyright issues, labeling costs/errors, and ethical biases.
****
Multi-Label Compound Expression Recognition: C-EXPR Database & Network
Dimitrios Kollias; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5589-5598
Research in automatic analysis of facial expressions mainly focuses on recognising the seven basic ones. However, compound expressions are more diverse and represent the complexity and subtlety of our daily affective displays more accurately. Limited research has been conducted for compound expression recognition (CER), because only a few databases exist, which are small, lab controlled, imbalanced and static. In this paper we present an in-the-wild A/V database, C-EXPR-DB, consisting of 400 videos of 200K frames, annotated in terms of 13 compound expressions, valence-arousal emotion descriptors, action units, speech, facial landmarks and attributes. We also propose C-EXPR-NET, a multi-task learning (MTL) method for CER and AU detection (AU-D); the latter task is introduced to enhance CER performance. For AU-D we incorporate AU semantic description along with visual information. For CER we use a multi-label formulation and the KL-divergence loss. We also propose a distribution matching loss for coupling CER and AU-D tasks to boost their performance and alleviate negative transfer (i.e., when MT model's performance is worse than that of at least one single-task model). An extensive experimental study has been conducted illustrating the excellent performance of C-EXPR-NET, validating the theoretical claims. Finally, C-EXPR-NET is shown to effectively generalize its knowledge in new emotion recognition contexts, in a zero-shot manner.
****
Masked Autoencoding Does Not Help Natural Language Supervision at Scale
Floris Weers, Vaishaal Shankar, Angelos Katharopoulos, Yinfei Yang, Tom Gunter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23432-23444
Self supervision and natural language supervision have emerged as two exciting ways to train general purpose image encoders which excel at a variety of downstream tasks. Recent works such as M3AE (Geng et al 2022) and SLIP (Mu et al 2022) have suggested that these approaches can be effectively combined, but most notably their results use small (<20M examples) pre-training datasets and don't effectively reflect the large-scale regime (>100M samples) that is commonly used for these approaches. Here we investigate whether a similar approach can be effective when trained with a much larger amount of data. We find that a combination of two state of the art approaches: masked auto-encoders, MAE (He et al 2021) and contrastive language image pre-training, CLIP (Radford et al 2021) provides a benefit over CLIP when trained on a corpus of 11.3M image-text pairs, but little to no benefit (as evaluated on a suite of common vision tasks) over CLIP when trained on a large corpus of 1.4B images. Our work provides some much needed clarity into the effectiveness (or lack thereof) of self supervision for large-scale image-text training.
****
CORA: Adapting CLIP for Open-Vocabulary Detection With Region Prompting and Anchor Pre-Matching
Xiaoshi Wu, Feng Zhu, Rui Zhao, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7031-7040

Open-vocabulary detection (OVD) is an object detection task aiming at detecting objects from novel categories beyond the base categories on which the detector is trained. Recent OVD methods rely on large-scale visual-language pre-trained models, such as CLIP, for recognizing novel objects. We identify the two core obstacles that need to be tackled when incorporating these models into detector training: (1) the distribution mismatch that happens when applying a VL-model trained on whole images to region recognition tasks; (2) the difficulty of localizing objects of unseen classes. To overcome these obstacles, we propose CORA, a DETR-style framework that adapts CLIP for Open-vocabulary detection by Region prompting and Anchor pre-matching. Region prompting mitigates the whole-to-region distribution gap by prompting the region features of the CLIP-based region classifier. Anchor pre-matching helps learning generalizable object localization by a class-aware matching mechanism. We evaluate CORA on the COCO OVD benchmark, where we achieve 41.7 AP50 on novel classes, which outperforms the previous SOTA by 2.4 AP50 even without resorting to extra training data. When extra training data is available, we train CORA+ on both ground-truth base-category annotations and additional pseudo bounding box labels computed by CORA. CORA+ achieves 43.1 AP50 on the COCO OVD benchmark and 28.1 box APr on the LVIS OVD benchmark. The code is available at https://github.com/tgxs002/CORA.
****
3DAvatarGAN: Bridging Domains for Personalized Editable Avatars
Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4552-4562
Modern 3D-GANs synthesize geometry and texture by training on large-scale datasets with a consistent structure. Training such models on stylized, artistic data, with often unknown, highly variable geometry, and camera information has not yet been shown possible. Can we train a 3D GAN on such artistic data, while maintaining multi-view consistency and texture quality? To this end, we propose an adaptation framework, where the source domain is a pre-trained 3D-GAN, while the target domain is a 2D-GAN trained on artistic datasets. We, then, distill the knowledge from a 2D generator to the source 3D generator. To do that, we first propose an optimization-based method to align the distributions of camera parameters across domains. Second, we propose regularizations necessary to learn high-quality texture, while avoiding degenerate geometric solutions, such as flat shapes. Third, we show a deformation-based technique for modeling exaggerated geometry of artistic domains, enabling---as a byproduct---personalized geometric editing. Finally, we propose a novel inversion method for 3D-GANs linking the latent spaces of the source and the target domains. Our contributions---for the first time---allow for the generation, editing, and animation of personalized artistic 3D avatars on artistic datasets.
****
Physics-Driven Diffusion Models for Impact Sound Synthesis From Videos
Kun Su, Kaizhi Qian, Eli Shlizerman, Antonio Torralba, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9749-9759
Modeling sounds emitted from physical object interactions is critical for immersive perceptual experiences in real and virtual worlds. Traditional methods of impact sound synthesis use physics simulation to obtain a set of physics parameters that could represent and synthesize the sound. However, they require fine details of both the object geometries and impact locations, which are rarely available in the real world and can not be applied to synthesize impact sounds from common videos. On the other hand, existing video-driven deep learning-based approaches could only capture the weak correspondence between visual content and impact sounds since they lack of physics knowledge. In this work, we propose a physics-driven diffusion model that can synthesize high-fidelity impact sound for a silent video clip. In addition to the video content, we propose to use additional physics priors to guide the impact sound synthesis procedure. The physics priors include both physics parameters that are directly estimated from noisy real-world impact sound examples without sophisticated setup and learned residual paramet

ers that interpret the sound environment via neural networks. We further impleme nt a novel diffusion model with specific training and inference strategies to co mbine physics priors and visual information for impact sound synthesis. Experime ntal results show that our model outperforms several existing systems in generat ing realistic impact sounds. More importantly, the physics-based representations are fully interpretable and transparent, thus enabling us to perform sound edit ing flexibly. We encourage the readers to visit our project page to watch demo v ideos with audio turned on to experience the results.
****
Transductive Few-Shot Learning With Prototype-Based Label Propagation by Iterati ve Graph Refinement
Hao Zhu, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2023, pp. 23996-24006
Few-shot learning (FSL) is popular due to its ability to adapt to novel classes. Compared with inductive few-shot learning, transductive models typically perfor m better as they leverage all samples of the query set. The two existing classes of methods, prototype-based and graph-based, have the disadvantages of inaccura te prototype estimation and sub-optimal graph construction with kernel functions , respectively. %, which hurt the performance. In this paper, we propose a novel prototype-based label propagation to solve these issues. Specifically, our grap h construction is based on the relation between prototypes and samples rather th an between samples. As prototypes are being updated, the graph changes.We also e stimate the label of each prototype instead of considering a prototype be the cl ass centre. On mini-ImageNet, tiered-ImageNet, CIFAR-FS and CUB datasets, we sho w the proposed method outperforms other state-of-the-art methods in transductive FSL and semi-supervised FSL when some unlabeled data accompanies the novel few-shot task.
****
Discriminative Co-Saliency and Background Mining Transformer for Co-Salient Obje ct Detection
Long Li, Junwei Han, Ni Zhang, Nian Liu, Salman Khan, Hisham Cholakkal, Rao Muha mmad Anwer, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Comput er Vision and Pattern Recognition (CVPR), 2023, pp. 7247-7256
Most previous co-salient object detection works mainly focus on extracting co-sa lient cues via mining the consistency relations across images while ignoring the explicit exploration of background regions. In this paper, we propose a Discrim inative co-saliency and background Mining Transformer framework (DMT) based on s everal economical multi-grained correlation modules to explicitly mine both co-s aliency and background information and effectively model their discrimination. S pecifically, we first propose region-to-region correlation modules to economical ly model inter-image relations for pixel-wise segmentation features. Then, we us e two types of predefined tokens to mine co-saliency and background information via our proposed contrast-induced pixel-to-token and co-saliency token-to-token correlation modules. We also design a token-guided feature refinement module to enhance the discriminability of the segmentation features under the guidance of the learned tokens. We perform iterative mutual promotion for the segmentation f eature extraction and token construction. Experimental results on three benchmar k datasets demonstrate the effectiveness of our proposed method. The source code is available at: https://github.com/dragonlee258079/DMT.
****
Alias-Free Convnets: Fractional Shift Invariance via Polynomial Activations
Hagay Michaeli, Tomer Michaeli, Daniel Soudry; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16333-16342
Although CNNs are believed to be invariant to translations, recent works have sh own this is not the case due to aliasing effects that stem from down-sampling la yers. The existing architectural solutions to prevent the aliasing effects are p artial since they do not solve those effects that originate in non-linearities. We propose an extended anti-aliasing method that tackles both down-sampling and non-linear layers, thus creating truly alias-free, shift-invariant CNNs. We show that the presented model is invariant to integer as well as fractional (i.e., s

ub-pixel) translations, thus outperforming other shift-invariant methods in terms of robustness to adversarial translations.
****
Binary Latent Diffusion
Ze Wang, Jiang Wang, Zicheng Liu, Qiang Qiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22576-22585
In this paper, we show that a binary latent space can be explored for compact yet expressive image representations. We model the bi-directional mappings between an image and the corresponding latent binary representation by training an auto-encoder with a Bernoulli encoding distribution. On the one hand, the binary latent space provides a compact discrete image representation of which the distribution can be modeled more efficiently than pixels or continuous latent representations. On the other hand, we now represent each image patch as a binary vector instead of an index of a learned cookbook as in discrete image representations with vector quantization. In this way, we obtain binary latent representations that allow for better image quality and high-resolution image representations without any multi-stage hierarchy in the latent space. In this binary latent space, images can now be generated effectively using a binary latent diffusion model tailored specifically for modeling the prior over the binary image representations. We present both conditional and unconditional image generation experiments with multiple datasets, and show that the proposed method performs comparably to state-of-the-art methods while dramatically improving the sampling efficiency to as few as 16 steps without using any test-time acceleration. The proposed framework can also be seamlessly scaled to 1024 x 1024 high-resolution image generation without resorting to latent hierarchy or multi-stage refinements.
****
Person Image Synthesis via Denoising Diffusion Model
Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5968-5976
The pose-guided person image generation task requires synthesizing photorealistic images of humans in arbitrary poses. The existing approaches use generative adversarial networks that do not necessarily maintain realistic textures or need dense correspondences that struggle to handle complex deformations and severe occlusions. In this work, we show how denoising diffusion models can be applied for high-fidelity person image synthesis with strong sample diversity and enhanced mode coverage of the learnt data distribution. Our proposed Person Image Diffusion Model (PIDM) disintegrates the complex transfer problem into a series of simpler forward-backward denoising steps. This helps in learning plausible source-to-target transformation trajectories that result in faithful textures and undistorted appearance details. We introduce a 'texture diffusion module' based on cross-attention to accurately model the correspondences between appearance and pose information available in source and target images. Further, we propose 'disentangled classifier-free guidance' to ensure close resemblance between the conditional inputs and the synthesized output in terms of both pose and appearance information. Our extensive results on two large-scale benchmarks and a user study demonstrate the photorealism of our proposed approach under challenging scenarios. We also show how our generated images can help in downstream tasks.
****
Shortcomings of Top-Down Randomization-Based Sanity Checks for Evaluations of Deep Neural Network Explanations
Alexander Binder, Leander Weber, Sebastian Lapuschkin, Grégoire Montavon, Klaus-Robert Müller, Wojciech Samek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16143-16152
While the evaluation of explanations is an important step towards trustworthy models, it needs to be done carefully, and the employed metrics need to be well-understood. Specifically model randomization testing can be overinterpreted if regarded as a primary criterion for selecting or discarding explanation methods. To address shortcomings of this test, we start by observing an experimental gap in the ranking of explanation methods between randomization-based sanity checks [1

] and model output faithfulness measures (e.g. [20]). We identify limitations of
 model-randomization-based sanity checks for the purpose of evaluating explanati
ons. Firstly, we show that uninformative attribution maps created with zero pixe
l-wise covariance easily achieve high scores in this type of checks. Secondly, w
e show that top-down model randomization preserves scales of forward pass activa
tions with high probability. That is, channels with large activations have a hig
h probility to contribute strongly to the output, even after randomization of th
e network on top of them. Hence, explanations after randomization can only be ex
pected to differ to a certain extent. This explains the observed experimental ga
p. In summary, these results demonstrate the inadequacy of model-randomization-b
ased sanity checks as a criterion to rank attribution methods.
****
Neural Part Priors: Learning To Optimize Part-Based Object Completion in RGB-D S
cans
Aleksei Bokhovkin, Angela Dai; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2023, pp. 9032-9042
3D scene understanding has seen significant advances in recent years, but has la
rgely focused on object understanding in 3D scenes with independent per-object p
redictions. We thus propose to learn Neural Part Priors (NPPs), parametric space
s of objects and their parts, that enable optimizing to fit to a new input 3D sc
an geometry with global scene consistency constraints. The rich structure of our
 NPPs enables accurate, holistic scene reconstruction across similar objects in
the scene. Both objects and their part geometries are characterized by coordinat
e field MLPs, facilitating optimization at test time to fit to input geometric o
bservations as well as similar objects in the input scan. This enables more accu
rate reconstructions than independent per-object predictions as a single forward
 pass, while establishing global consistency within a scene. Experiments on the
ScanNet dataset demonstrate that NPPs significantly outperforms the state-of-the
-art in part decomposition and object completion in real-world scenes.
****
Adaptive Assignment for Geometry Aware Local Feature Matching
Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding,
 Fan Tang, Chengjie Wang; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2023, pp. 5425-5434
The detector-free feature matching approaches are currently attracting great att
ention thanks to their excellent performance. However, these methods still strug
gle at large-scale and viewpoint variations, due to the geometric inconsistency
resulting from the application of the mutual nearest neighbour criterion (i.e.,
one-to-one assignment) in patch-level matching. Accordingly, we introduce AdaMat
cher, which first accomplishes the feature correlation and co-visible area estim
ation through an elaborate feature interaction module, then performs adaptive as
signment on patch-level matching while estimating the scales between images, and
 finally refines the co-visible matches through scale alignment and sub-pixel re
gression module. Extensive experiments show that AdaMatcher outperforms solid ba
selines and achieves state-of-the-art results on many downstream tasks. Addition
ally, the adaptive assignment and sub-pixel refinement module can be used as a r
efinement network for other matching methods, such as SuperGlue, to boost their
performance further. The code will be publicly available at https://github.com/A
byssGaze/AdaMatcher.
****
Initialization Noise in Image Gradients and Saliency Maps
Ann-Christin Woerl, Jan Disselhoff, Michael Wand; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1766-1775
In this paper, we examine gradients of logits of image classification CNNs by in
put pixel values. We observe that these fluctuate considerably with training ran
domness, such as the random initialization of the networks. We extend our study
to gradients of intermediate layers, obtained via GradCAM, as well as popular ne
twork saliency estimators such as DeepLIFT, SHAP, LIME, Integrated Gradients, an
d SmoothGrad. While empirical noise levels vary, qualitatively different attribu
tions to image features are still possible with all of these, which comes with i

mplications for interpreting such attributions, in particular when seeking data-driven explanations of the phenomenon generating the data. Finally, we demonstrate that the observed artefacts can be removed by marginalization over the initialization distribution by simple stochastic integration.
****
FLAG3D: A 3D Fitness Activity Dataset With Language Instruction
Yansong Tang, Jinpeng Liu, Aoyang Liu, Bin Yang, Wenxun Dai, Yongming Rao, Jiwen Lu, Jie Zhou, Xiu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22106-22117
With the continuously thriving popularity around the world, fitness activity analytic has become an emerging research topic in computer vision. While a variety of new tasks and algorithms have been proposed recently, there are growing hunger for data resources involved in high-quality data, fine-grained labels, and diverse environments. In this paper, we present FLAG3D, a large-scale 3D fitness activity dataset with language instruction containing 180K sequences of 60 categories. FLAG3D features the following three aspects: 1) accurate and dense 3D human pose captured from advanced MoCap system to handle the complex activity and large movement, 2) detailed and professional language instruction to describe how to perform a specific activity, 3) versatile video resources from a high-tech MoCap system, rendering software, and cost-effective smartphones in natural environments. Extensive experiments and in-depth analysis show that FLAG3D contributes great research value for various challenges, such as cross-domain human action recognition, dynamic human mesh recovery, and language-guided human action generation. Our dataset and source code are publicly available at https://andytang15.github.io/FLAG3D.
****
Implicit Neural Head Synthesis via Controllable Local Deformation Fields
Chuhan Chen, Matthew O’Toole, Gaurav Bharaj, Pablo Garrido; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 416-426
High-quality reconstruction of controllable 3D head avatars from 2D videos is highly desirable for virtual human applications in movies, games, and telepresence. Neural implicit fields provide a powerful representation to model 3D head avatars with personalized shape, expressions, and facial parts, e.g., hair and mouth interior, that go beyond the linear 3D morphable model (3DMM). However, existing methods do not model faces with fine-scale facial features, or local control of facial parts that extrapolate asymmetric expressions from monocular videos. Further, most condition only on 3DMM parameters with poor(er) locality, and resolve local features with a global neural field. We build on part-based implicit shape models that decompose a global deformation field into local ones. Our novel formulation models multiple implicit deformation fields with local semantic rig-like control via 3DMM-based parameters, and representative facial landmarks. Further, we propose a local control loss and attention mask mechanism that promote sparsity of each learned deformation field. Our formulation renders sharper locally controllable nonlinear deformations than previous implicit monocular approaches, especially mouth interior, asymmetric expressions, and facial details. Project page:https://imaging.cs.cmu.edu/local_deformation_fields/
****
NeuralUDF: Learning Unsigned Distance Fields for Multi-View Reconstruction of Surfaces With Arbitrary Topologies
Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20834-20843
We present a novel method, called NeuralUDF, for reconstructing surfaces with arbitrary topologies from 2D images via volume rendering. Recent advances in neural rendering based reconstruction have achieved compelling results. However, these methods are limited to objects with closed surfaces since they adopt Signed Distance Function (SDF) as surface representation which requires the target shape to be divided into inside and outside. In this paper, we propose to represent surfaces as the Unsigned Distance Function (UDF) and develop a new volume renderin

g scheme to learn the neural UDF representation. Specifically, a new density function that correlates the property of UDF with the volume rendering scheme is introduced for robust optimization of the UDF fields. Experiments on the DTU and DeepFashion3D datasets show that our method not only enables high-quality reconstruction of non-closed shapes with complex typologies, but also achieves comparable performance to the SDF based methods on the reconstruction of closed surfaces. Visit our project page at https://www.xxlong.site/NeuralUDF/.
****
Towards Trustable Skin Cancer Diagnosis via Rewriting Model's Decision
Siyuan Yan, Zhen Yu, Xuelin Zhang, Dwarikanath Mahapatra, Shekhar S. Chandra, Monika Janda, Peter Soyer, Zongyuan Ge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11568-11577
Deep neural networks have demonstrated promising performance on image recognition tasks. However, they may heavily rely on confounding factors, using irrelevant artifacts or bias within the dataset as the cue to improve performance. When a model performs decision-making based on these spurious correlations, it can become untrustable and lead to catastrophic outcomes when deployed in the real-world scene. In this paper, we explore and try to solve this problem in the context of skin cancer diagnosis. We introduce a human-in-the-loop framework in the model training process such that users can observe and correct the model's decision logic when confounding behaviors happen. Specifically, our method can automatically discover confounding factors by analyzing the co-occurrence behavior of the samples. It is capable of learning confounding concepts using easily obtained concept exemplars. By mapping the blackbox model's feature representation onto an explainable concept space, human users can interpret the concept and intervene via a first order-logic instruction. We systematically evaluate our method on our newly crafted, well-controlled skin lesion dataset and several public skin lesion datasets. Experiments show that our method can effectively detect and remove confounding factors from datasets without any prior knowledge about the category distribution and does not require fully annotated concept labels. We also show that our method enables the model to focus on clinicalrelated concepts, improving the model's performance and trustworthiness during model inference.
****
Curricular Object Manipulation in LiDAR-Based Object Detection
Ziyue Zhu, Qiang Meng, Xiao Wang, Ke Wang, Liujiang Yan, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1125-1135
This paper explores the potential of curriculum learning in LiDAR-based 3D object detection by proposing a curricular object manipulation (COM) framework. The framework embeds the curricular training strategy into both the loss design and the augmentation process. For the loss design, we propose the COMLoss to dynamically predict object-level difficulties and emphasize objects of different difficulties based on training stages. On top of the widely-used augmentation technique called GT-Aug in LiDAR detection tasks, we propose a novel COMAug strategy which first clusters objects in ground-truth database based on well-designed heuristics. Group-level difficulties rather than individual ones are then predicted and updated during training for stable results. Model performance and generalization capabilities can be improved by sampling and augmenting progressively more difficult objects into the training points. Extensive experiments and ablation studies reveal the superior and generality of the proposed framework. The code is available at https://github.com/ZZY816/COM.
****
Collaborative Static and Dynamic Vision-Language Streams for Spatio-Temporal Video Grounding
Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, Wei-Shi Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23100-23109
Spatio-Temporal Video Grounding (STVG) aims to localize the target object spatially and temporally according to the given language query. It is a challenging task in which the model should well understand dynamic visual cues (e.g., motions)

and static visual cues (e.g., object appearances) in the language description, which requires effective joint modeling of spatio-temporal visual-linguistic dependencies. In this work, we propose a novel framework in which a static vision-language stream and a dynamic vision-language stream are developed to collaboratively reason the target tube. The static stream performs cross-modal understanding in a single frame and learns to attend to the target object spatially according to intra-frame visual cues like object appearances. The dynamic stream models visual-linguistic dependencies across multiple consecutive frames to capture dynamic cues like motions. We further design a novel cross-stream collaborative block between the two streams, which enables the static and dynamic streams to transfer useful and complementary information from each other to achieve collaborative reasoning. Experimental results show the effectiveness of the collaboration of the two streams and our overall framework achieves new state-of-the-art performance on both HCSTVG and VidSTG datasets.
****
Shape-Constraint Recurrent Flow for 6D Object Pose Estimation
Yang Hai, Rui Song, Jiaojiao Li, Yinlin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4831-4840
Most recent 6D object pose estimation methods rely on 2D optical flow networks to refine their results. However, these optical flow methods typically do not consider any 3D shape information of the targets during matching, making them suffer in 6D object pose estimation. In this work, we propose a shape-constraint recurrent flow network for 6D object pose estimation, which embeds the 3D shape information of the targets into the matching procedure. We first introduce a flow-to-pose component to learn an intermediate pose from the current flow estimation, then impose a shape constraint from the current pose on the lookup space of the 4D correlation volume for flow estimation, which reduces the matching space significantly and is much easier to learn. Finally, we optimize the flow and pose simultaneously in a recurrent manner until convergence. We evaluate our method on three challenging 6D object pose datasets and show that it outperforms the state of the art in both accuracy and efficiency.
****
FeatER: An Efficient Network for Human Reconstruction via Feature Map-Based TransformER
Ce Zheng, Matias Mendieta, Taojiannan Yang, Guo-Jun Qi, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13945-13954
Recently, vision transformers have shown great success in a set of human reconstruction tasks such as 2D human pose estimation (2D HPE), 3D human pose estimation (3D HPE), and human mesh reconstruction (HMR) tasks. In these tasks, feature map representations of the human structural information are often extracted first from the image by a CNN (such as HRNet), and then further processed by transformer to predict the heatmaps (encodes each joint's location into a feature map with a Gaussian distribution) for HPE or HMR. However, existing transformer architectures are not able to process these feature map inputs directly, forcing an unnatural flattening of the location-sensitive human structural information. Furthermore, much of the performance benefit in recent HPE and HMR methods has come at the cost of ever-increasing computation and memory needs. Therefore, to simultaneously address these problems, we propose FeatER, a novel transformer design which preserves the inherent structure of feature map representations when modeling attention while reducing the memory and computational costs. Taking advantage of FeatER, we build an efficient network for a set of human reconstruction tasks including 2D HPE, 3D HPE, and HMR. A feature map reconstruction module is applied to improve the performance of the estimated human pose and mesh. Extensive experiments demonstrate the effectiveness of FeatER on various human pose and mesh datasets. For instance, FeatER outperforms the SOTA method MeshGraphormer by requiring 5% of Params (total parameters) and 16% of MACs (the Multiply-Accumulate Operations) on Human3.6M and 3DPW datasets. Code will be publicly available.
****
Micron-BERT: BERT-Based Facial Micro-Expression Recognition

Xuan-Bac Nguyen, Chi Nhan Duong, Xin Li, Susan Gauch, Han-Seok Seo, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1482-1492
Micro-expression recognition is one of the most challenging topics in affective computing. It aims to recognize tiny facial movements difficult for humans to perceive in a brief period, i.e., 0.25 to 0.5 seconds. Recent advances in pre-training deep Bidirectional Transformers (BERT) have significantly improved self-supervised learning tasks in computer vision. However, the standard BERT in vision problems is designed to learn only from full images or videos, and the architecture cannot accurately detect details of facial micro-expressions. This paper presents Micron-BERT (u-BERT), a novel approach to facial micro-expression recognition. The proposed method can automatically capture these movements in an unsupervised manner based on two key ideas. First, we employ Diagonal Micro-Attention (DMA) to detect tiny differences between two frames. Second, we introduce a new Patch of Interest (PoI) module to localize and highlight micro-expression interest regions and simultaneously reduce noisy backgrounds and distractions. By incorporating these components into an end-to-end deep network, the proposed u-BERT significantly outperforms all previous work in various micro-expression tasks. u-BERT can be trained on a large-scale unlabeled dataset, i.e., up to 8 million images, and achieves high accuracy on new unseen facial micro-expression datasets. Empirical experiments show u-BERT consistently outperforms state-of-the-art performance on four micro-expression benchmarks, including SAMM, CASME II, SMIC, and CASME3, by significant margins. Code will be available at https://github.com/uark-cviu/Micron-BERT
****
Residual Degradation Learning Unfolding Framework With Mixing Priors Across Spectral and Spatial for Compressive Spectral Imaging
Yubo Dong, Dahua Gao, Tian Qiu, Yuyan Li, Minxi Yang, Guangming Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22262-22271
To acquire a snapshot spectral image, coded aperture snapshot spectral imaging (CASSI) is proposed. A core problem of the CASSI system is to recover the reliable and fine underlying 3D spectral cube from the 2D measurement. By alternately solving a data subproblem and a prior subproblem, deep unfolding methods achieve good performance. However, in the data subproblem, the used sensing matrix is ill-suited for the real degradation process due to the device errors caused by phase aberration, distortion; in the prior subproblem, it is important to design a suitable model to jointly exploit both spatial and spectral priors. In this paper, we propose a Residual Degradation Learning Unfolding Framework (RDLUF), which bridges the gap between the sensing matrix and the degradation process. Moreover, a MixS2 Transformer is designed via mixing priors across spectral and spatial to strengthen the spectral-spatial representation capability. Finally, plugging the MixS2 Transformer into the RDLUF leads to an end-to-end trainable and interpretable neural network RDLUF-MixS2. Experimental results establish the superior performance of the proposed method over existing ones.
****
Visibility Constrained Wide-Band Illumination Spectrum Design for Seeing-in-the-Dark
Muyao Niu, Zhuoxiao Li, Zhihang Zhong, Yinqiang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13976-13985
Seeing-in-the-dark is one of the most important and challenging computer vision tasks due to its wide applications and extreme complexities of in-the-wild scenarios. Existing arts can be mainly divided into two threads: 1) RGB-dependent methods restore information using degraded RGB inputs only (e.g., low-light enhancement), 2) RGB-independent methods translate images captured under auxiliary near-infrared (NIR) illuminants into RGB domain (e.g., NIR2RGB translation). The latter is very attractive since it works in complete darkness and the illuminants are visually friendly to naked eyes, but tends to be unstable due to its intrinsic ambiguities. In this paper, we try to robustify NIR2RGB translation by designi

ng the optimal spectrum of auxiliary illumination in the wide-band VIS-NIR range, while keeping visual friendliness. Our core idea is to quantify the visibility constraint implied by the human vision system and incorporate it into the design pipeline. By modeling the formation process of images in the VIS-NIR range, the optimal multiplexing of a wide range of LEDs is automatically designed in a fully differentiable manner, within the feasible region defined by the visibility constraint. We also collect a substantially expanded VIS-NIR hyperspectral image dataset for experiments by using a customized 50-band filter wheel. Experimental results show that the task can be significantly improved by using the optimized wide-band illumination than using NIR only. Codes Available: https://github.com/MyNiuuu/VCSD.
****
PanelNet: Understanding 360 Indoor Environment via Panel Representation
Haozheng Yu, Lu He, Bing Jian, Weiwei Feng, Shan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 878-887
Indoor 360 panoramas have two essential properties. (1) The panoramas are continuous and seamless in the horizontal direction. (2) Gravity plays an important role in indoor environment design. By leveraging these properties, we present PanelNet, a framework that understands indoor environments using a novel panel representation of 360 images. We represent an equirectangular projection (ERP) as consecutive vertical panels with corresponding 3D panel geometry. To reduce the negative impact of panoramic distortion, we incorporate a panel geometry embedding network that encodes both the local and global geometric features of a panel. To capture the geometric context in room design, we introduce Local2Global Transformer, which aggregates local information within a panel and panel-wise global context. It greatly improves the model performance with low training overhead. Our method outperforms existing methods on indoor 360 depth estimation and shows competitive results against state-of-the-art approaches on the task of indoor layout estimation and semantic segmentation.
****
Learning With Noisy Labels via Self-Supervised Adversarial Noisy Masking
Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Cai Rong Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16186-16195
Collecting large-scale datasets is crucial for training deep models, annotating the data, however, inevitably yields noisy labels, which poses challenges to deep learning algorithms. Previous efforts tend to mitigate this problem via identifying and removing noisy samples or correcting their labels according to the statistical properties (e.g., loss values) among training samples. In this paper, we aim to tackle this problem from a new perspective, delving into the deep feature maps, we empirically find that models trained with clean and mislabeled samples manifest distinguishable activation feature distributions. From this observation, a novel robust training approach termed adversarial noisy masking is proposed. The idea is to regularize deep features with a label quality guided masking scheme, which adaptively modulates the input data and label simultaneously, preventing the model to overfit noisy samples. Further, an auxiliary task is designed to reconstruct input data, it naturally provides noise-free self-supervised signals to reinforce the generalization ability of deep models. The proposed method is simple and flexible, it is tested on both synthetic and real-world noisy datasets, where significant improvements are achieved over previous state-of-the-art methods.
****
PoseExaminer: Automated Testing of Out-of-Distribution Robustness in Human Pose and Shape Estimation
Qihao Liu, Adam Kortylewski, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 672-681
Human pose and shape (HPS) estimation methods achieve remarkable results. However, current HPS benchmarks are mostly designed to test models in scenarios that are similar to the training data. This can lead to critical situations in real-wo

rld applications when the observed data differs significantly from the training data and hence is out-of-distribution (OOD). It is therefore important to test and improve the OOD robustness of HPS methods. To address this fundamental problem, we develop a simulator that can be controlled in a fine-grained manner using interpretable parameters to explore the manifold of images of human pose, e.g. by varying poses, shapes, and clothes. We introduce a learning-based testing method, termed PoseExaminer, that automatically diagnoses HPS algorithms by searching over the parameter space of human pose images to find the failure modes. Our strategy for exploring this high-dimensional parameter space is a multi-agent reinforcement learning system, in which the agents collaborate to explore different parts of the parameter space. We show that our PoseExaminer discovers a variety of limitations in current state-of-the-art models that are relevant in real-world scenarios but are missed by current benchmarks. For example, it finds large regions of realistic human poses that are not predicted correctly, as well as reduced performance for humans with skinny and corpulent body shapes. In addition, we show that fine-tuning HPS methods by exploiting the failure modes found by PoseExaminer improve their robustness and even their performance on standard benchmarks by a significant margin. The code are available for research purposes.
****
GamutMLP: A Lightweight MLP for Color Loss Recovery
Hoang M. Le, Brian Price, Scott Cohen, Michael S. Brown; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18268-18277
Cameras and image-editing software often process images in the wide-gamut ProPhoto color space, encompassing 90% of all visible colors. However, when images are encoded for sharing, this color-rich representation is transformed and clipped to fit within the small-gamut standard RGB (sRGB) color space, representing only 30% of visible colors. Recovering the lost color information is challenging due to the clipping procedure. Inspired by neural implicit representations for 2D images, we propose a method that optimizes a lightweight multi-layer-perceptron (MLP) model during the gamut reduction step to predict the clipped values. GamutMLP takes approximately 2 seconds to optimize and requires only 23 KB of storage. The small memory footprint allows our GamutMLP model to be saved as metadata in the sRGB image---the model can be extracted when needed to restore wide-gamut color values. We demonstrate the effectiveness of our approach for color recovery and compare it with alternative strategies, including pre-trained DNN-based gamut expansion networks and other implicit neural representation methods. As part of this effort, we introduce a new color gamut dataset of 2200 wide-gamut/small-gamut images for training and testing.
****
Instance-Aware Domain Generalization for Face Anti-Spoofing
Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Xuequan Lu, Ran Yi, Shouhong Ding, Lizhuang Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20453-20463
Face anti-spoofing (FAS) based on domain generalization (DG) has been recently studied to improve the generalization on unseen scenarios. Previous methods typically rely on domain labels to align the distribution of each domain for learning domain-invariant representations. However, artificial domain labels are coarse-grained and subjective, which cannot reflect real domain distributions accurately. Besides, such domain-aware methods focus on domain-level alignment, which is not fine-grained enough to ensure that learned representations are insensitive to domain styles. To address these issues, we propose a novel perspective for DG FAS that aligns features on the instance level without the need for domain labels. Specifically, Instance-Aware Domain Generalization framework is proposed to learn the generalizable feature by weakening the features' sensitivity to instance-specific styles. Concretely, we propose Asymmetric Instance Adaptive Whitening to adaptively eliminate the style-sensitive feature correlation, boosting the generalization. Moreover, Dynamic Kernel Generator and Categorical Style Assembly are proposed to first extract the instance-specific features and then generate the style-diversified features with large style shifts, respectively, further fa

cilitating the learning of style-insensitive features. Extensive experiments and analysis demonstrate the superiority of our method over state-of-the-art competitors. Code will be publicly available at this link: https://github.com/qianyuzqy/IADG.
****
GANHead: Towards Generative Animatable Neural Head Avatars
Sijing Wu, Yichao Yan, Yunhao Li, Yuhao Cheng, Wenhan Zhu, Ke Gao, Xiaobo Li, Guangtao Zhai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 437-447
To bring digital avatars into people's lives, it is highly demanded to efficiently generate complete, realistic, and animatable head avatars. This task is challenging, and it is difficult for existing methods to satisfy all the requirements at once. To achieve these goals, we propose GANHead (Generative Animatable Neural Head Avatar), a novel generative head model that takes advantages of both the fine-grained control over the explicit expression parameters and the realistic rendering results of implicit representations. Specifically, GANHead represents coarse geometry, fine-gained details and texture via three networks in canonical space to obtain the ability to generate complete and realistic head avatars. To achieve flexible animation, we define the deformation filed by standard linear blend skinning (LBS), with the learned continuous pose and expression bases and LBS weights. This allows the avatars to be directly animated by FLAME parameters and generalize well to unseen poses and expressions. Compared to state-of-the-art (SOTA) methods, GANHead achieves superior performance on head avatar generation and raw scan fitting.
****
Towards Domain Generalization for Multi-View 3D Object Detection in Bird-Eye-View
Shuo Wang, Xinhai Zhao, Hai-Ming Xu, Zehui Chen, Dameng Yu, Jiahao Chang, Zhen Yang, Feng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13333-13342
Multi-view 3D object detection (MV3D-Det) in Bird-Eye-View (BEV) has drawn extensive attention due to its low cost and high efficiency. Although new algorithms for camera-only 3D object detection have been continuously proposed, most of them may risk drastic performance degradation when the domain of input images differs from that of training. In this paper, we first analyze the causes of the domain gap for the MV3D-Det task. Based on the covariate shift assumption, we find that the gap mainly attributes to the feature distribution of BEV, which is determined by the quality of both depth estimation and 2D image's feature representation. To acquire a robust depth prediction, we propose to decouple the depth estimation from the intrinsic parameters of the camera (i.e. the focal length) through converting the prediction of metric depth to that of scale-invariant depth and perform dynamic perspective augmentation to increase the diversity of the extrinsic parameters (i.e. the camera poses) by utilizing homography. Moreover, we modify the focal length values to create multiple pseudo-domains and construct an adversarial training loss to encourage the feature representation to be more domain-agnostic. Without bells and whistles, our approach, namely DG-BEV, successfully alleviates the performance drop on the unseen target domain without impairing the accuracy of the source domain. Extensive experiments on Waymo, nuScenes, and Lyft, demonstrate the generalization and effectiveness of our approach.
****
Robust and Scalable Gaussian Process Regression and Its Applications
Yifan Lu, Jiayi Ma, Leyuan Fang, Xin Tian, Junjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21950-21959
This paper introduces a robust and scalable Gaussian process regression (GPR) model via variational learning. This enables the application of Gaussian processes to a wide range of real data, which are often large-scale and contaminated by outliers. Towards this end, we employ a mixture likelihood model where outliers are assumed to be sampled from a uniform distribution. We next derive a variational formulation that jointly infers the mode of data, i.e., inlier or outlier, as

well as hyperparameters by maximizing a lower bound of the true log marginal likelihood. Compared to previous robust GPR, our formulation approximates the exact posterior distribution. The inducing variable approximation and stochastic variational inference are further introduced to our variational framework, extending our model to large-scale data. We apply our model to two challenging real-world applications, namely feature matching and dense gene expression imputation. Extensive experiments demonstrate the superiority of our model in terms of robustness and speed. Notably, when matching 4k feature points, its inference is completed in milliseconds with almost no false matches. The code is at https://github.com/YifanLu2000/Robust-Scalable-GPR.
****
Deep Dive Into Gradients: Better Optimization for 3D Object Detection With Gradient-Corrected IoU Supervision
Qi Ming, Lingjuan Miao, Zhe Ma, Lin Zhao, Zhiqiang Zhou, Xuhui Huang, Yuanpei Chen, Yufei Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5136-5145
Intersection-over-Union (IoU) is the most popular metric to evaluate regression performance in 3D object detection. Recently, there are also some methods applying IoU to the optimization of 3D bounding box regression. However, we demonstrate through experiments and mathematical proof that the 3D IoU loss suffers from abnormal gradient w.r.t. angular error and object scale, which further leads to slow convergence and suboptimal regression process, respectively. In this paper, we propose a Gradient-Corrected IoU (GCIoU) loss to achieve fast and accurate 3D bounding box regression. Specifically, a gradient correction strategy is designed to endow 3D IoU loss with a reasonable gradient. It ensures that the model converges quickly in the early stage of training, and helps to achieve fine-grained refinement of bounding boxes in the later stage. To solve suboptimal regression of 3D IoU loss for objects at different scales, we introduce a gradient rescaling strategy to adaptively optimize the step size. Finally, we integrate GCIoU Loss into multiple models to achieve stable performance gains and faster model convergence. Experiments on KITTI dataset demonstrate superiority of the proposed method. The code is available at https://github.com/ming71/GCIoU-loss.
****
Doubly Right Object Recognition: A Why Prompt for Visual Rationales
Chengzhi Mao, Revant Teotia, Amrutha Sundar, Sachit Menon, Junfeng Yang, Xin Wang, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2722-2732
Many visual recognition models are evaluated only on their classification accuracy, a metric for which they obtain strong performance. In this paper, we investigate whether computer vision models can also provide correct rationales for their predictions. We propose a "doubly right" object recognition benchmark, where the metric requires the model to simultaneously produce both the right labels as well as the right rationales. We find that state-of-the-art visual models, such as CLIP, often provide incorrect rationales for their categorical predictions. However, by transferring the rationales from language models into visual representations through a tailored dataset, we show that we can learn a "why prompt," which adapts large visual representations to produce correct rationales. Visualizations and empirical experiments show that our prompts significantly improve performance on doubly right object recognition, in addition to zero-shot transfer to unseen tasks and datasets.
****
Shepherding Slots to Objects: Towards Stable and Robust Object-Centric Learning
Jinwoo Kim, Janghyuk Choi, Ho-Jin Choi, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19198-19207
Object-centric learning (OCL) aspires general and com- positional understanding of scenes by representing a scene as a collection of object-centric representations. OCL has also been extended to multi-view image and video datasets to apply various data-driven inductive biases by utilizing geometric or temporal information in the multi-image data. Single-view images carry less information about how

to disentangle a given scene than videos or multi-view im- ages do. Hence, owing to the difficulty of applying induc- tive biases, OCL for single-view images still remains chal- lenging, resulting in inconsistent learning of object-centric representation. To this end, we introduce a novel OCL framework for single-view images, SLot Attention via SHep- herding (SLASH), which consists of two simple-yet-effective modules on top of Slot Attention. The new modules, At- tention Refining Kernel (ARK) and Intermediate Point Pre- dictor and Encoder (IPPE), respectively, prevent slots from being distracted by the background noise and indicate lo- cations for slots to focus on to facilitate learning of object- centric representation. We also propose a weak- and semi- supervision approach for OCL, whilst our proposed frame- work can be used without any assistant annotation during the inference. Experiments show that our proposed method enables consistent learning of object-centric representa- tion and achieves strong performance across four datasets. Code is available at https://github.com/object- understanding/SLASH.
****
High-Fidelity Event-Radiance Recovery via Transient Event Frequency
Jin Han, Yuta Asano, Boxin Shi, Yinqiang Zheng, Imari Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20616-20625
High-fidelity radiance recovery plays a crucial role in scene information reconstruction and understanding. Conventional cameras suffer from limited sensitivity in dynamic range, bit depth, and spectral response, etc. In this paper, we propose to use event cameras with bio-inspired silicon sensors, which are sensitive to radiance changes, to recover precise radiance values. We reveal that, under active lighting conditions, the transient frequency of event signals triggering linearly reflects the radiance value. We propose an innovative method to convert the high temporal resolution of event signals into precise radiance values. The precise radiance values yields several capabilities in image analysis. We demonstrate the feasibility of recovering radiance values solely from the transient event frequency (TEF) through multiple experiments.
****
NeMo: Learning 3D Neural Motion Fields From Multiple Video Instances of the Same Action
Kuan-Chieh Wang, Zhenzhen Weng, Maria Xenochristou, João Pedro Araújo, Jeffrey Gu, Karen Liu, Serena Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22129-22138
The task of reconstructing 3D human motion has wide-ranging applications. The gold standard Motion capture (MoCap) systems are accurate but inaccessible to the general public due to their cost, hardware, and space constraints. In contrast, monocular human mesh recovery (HMR) methods are much more accessible than MoCap as they take single-view videos as inputs. Replacing the multi-view MoCap systems with a monocular HMR method would break the current barriers to collecting accurate 3D motion thus making exciting applications like motion analysis and motion-driven animation accessible to the general public. However, the performance of existing HMR methods degrades when the video contains challenging and dynamic motion that is not in existing MoCap datasets used for training. This reduces its appeal as dynamic motion is frequently the target in 3D motion recovery in the aforementioned applications. Our study aims to bridge the gap between monocular HMR and multi-view MoCap systems by leveraging information shared across multiple video instances of the same action. We introduce the Neural Motion (NeMo) field. It is optimized to represent the underlying 3D motions across a set of videos of the same action. Empirically, we show that NeMo can recover 3D motion in sports using videos from the Penn Action dataset, where NeMo outperforms existing HMR methods in terms of 2D keypoint detection. To further validate NeMo using 3D metrics, we collected a small MoCap dataset mimicking actions in Penn Action, and show that NeMo achieves better 3D reconstruction compared to various baselines.
****
RIATIG: Reliable and Imperceptible Adversarial Text-to-Image Generation With Nat

ural Prompts
Han Liu, Yuhao Wu, Shixuan Zhai, Bo Yuan, Ning Zhang; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20585-
20594
The field of text-to-image generation has made remarkable strides in creating hi
gh-fidelity and photorealistic images. As this technology gains popularity, ther
e is a growing concern about its potential security risks. However, there has be
en limited exploration into the robustness of these models from an adversarial p
erspective. Existing research has primarily focused on untargeted settings, and
lacks holistic consideration for reliability (attack success rate) and stealthin
ess (imperceptibility). In this paper, we propose RIATIG, a reliable and imperce
ptible adversarial attack against text-to-image models via inconspicuous example
s. By formulating the example crafting as an optimization process and solving it
 using a genetic-based method, our proposed attack can generate imperceptible pr
ompts for text-to-image generation models in a reliable way. Evaluation of six p
opular text-to-image generation models demonstrates the efficiency and stealthin
ess of our attack in both white-box and black-box settings. To allow the communi
ty to build on top of our findings, we've made the artifacts available.
****
Distilling Neural Fields for Real-Time Articulated Shape Reconstruction
Jeff Tan, Gengshan Yang, Deva Ramanan; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4692-4701
We present a method for reconstructing articulated 3D models from videos in real
-time, without test-time optimization or manual 3D supervision at training time.
 Prior work often relies on pre-built deformable models (e.g. SMAL/SMPL), or slo
w per-scene optimization through differentiable rendering (e.g. dynamic NeRFs).
Such methods fail to support arbitrary object categories, or are unsuitable for
real-time applications. To address the challenge of collecting large-scale 3D tr
aining data for arbitrary deformable object categories, our key insight is to us
e off-the-shelf video-based dynamic NeRFs as 3D supervision to train a fast feed
-forward network, turning 3D shape and motion prediction into a supervised disti
llation task. Our temporal-aware network uses articulated bones and blend skinni
ng to represent arbitrary deformations, and is self-supervised on video datasets
 without requiring 3D shapes or viewpoints as input. Through distillation, our n
etwork learns to 3D-reconstruct unseen articulated objects at interactive frame
rates. Our method yields higher-fidelity 3D reconstructions than prior real-time
 methods for animals, with the ability to render realistic images at novel viewp
oints and poses.
****
GLIGEN: Open-Set Grounded Text-to-Image Generation
Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Ch
unyuan Li, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2023, pp. 22511-22521
Large-scale text-to-image diffusion models have made amazing advances. However,
the status quo is to use text input alone, which can impede controllability. In
this work, we propose GLIGEN: Open-Set Grounded Text-to-Image Generation, a nove
l approach that builds upon and extends the functionality of existing pre-traine
d text-to-image diffusion models by enabling them to also be conditioned on grou
nding inputs. To preserve the vast concept knowledge of the pre-trained model, w
e freeze all of its weights and inject the grounding information into new traina
ble layers via a gated mechanism. Our model achieves open-world grounded text2im
g generation with caption and bounding box condition inputs, and the grounding a
bility generalizes well to novel spatial configurations and concepts. GLIGEN's z
ero-shot performance on COCO and LVIS outperforms existing supervised layout-to-
image baselines by a large margin.
****
Q: How To Specialize Large Vision-Language Models to Data-Scarce VQA Tasks? A: S
elf-Train on Unlabeled Images!
Zaid Khan, Vijay Kumar BG, Samuel Schulter, Xiang Yu, Yun Fu, Manmohan Chandrake
r; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni

tion (CVPR), 2023, pp. 15005-15015
Finetuning a large vision language model (VLM) on a target dataset after large scale pretraining is a dominant paradigm in visual question answering (VQA). Datasets for specialized tasks such as knowledge-based VQA or VQA in non natural-image domains are orders of magnitude smaller than those for general-purpose VQA. While collecting additional labels for specialized tasks or domains can be challenging, unlabeled images are often available. We introduce SelTDA (Self-Taught Data Augmentation), a strategy for finetuning large VLMs on small-scale VQA datasets. SelTDA uses the VLM and target dataset to build a teacher model that can generate question-answer pseudolabels directly conditioned on an image alone, allowing us to pseudolabel unlabeled images. SelTDA then finetunes the initial VLM on the original dataset augmented with freshly pseudolabeled images. We describe a series of experiments showing that our self-taught data augmentation increases robustness to adversarially searched questions, counterfactual examples, and rephrasings, it improves domain generalization, and results in greater retention of numerical reasoning skills. The proposed strategy requires no additional annotations or architectural modifications, and is compatible with any modern encoder-decoder multimodal transformer. Code available at https://github.com/codezakh/SelTDA
****
IPCC-TP: Utilizing Incremental Pearson Correlation Coefficient for Joint Multi-Agent Trajectory Prediction
Dekai Zhu, Guangyao Zhai, Yan Di, Fabian Manhardt, Hendrik Berkemeyer, Tuan Tran, Nassir Navab, Federico Tombari, Benjamin Busam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5507-5516
Reliable multi-agent trajectory prediction is crucial for the safe planning and control of autonomous systems. Compared with single-agent cases, the major challenge in simultaneously processing multiple agents lies in modeling complex social interactions caused by various driving intentions and road conditions. Previous methods typically leverage graph-based message propagation or attention mechanism to encapsulate such interactions in the format of marginal probabilistic distributions. However, it is inherently sub-optimal. In this paper, we propose IPCC-TP, a novel relevance-aware module based on Incremental Pearson Correlation Coefficient to improve multi-agent interaction modeling. IPCC-TP learns pairwise joint Gaussian Distributions through the tightly-coupled estimation of the means and covariances according to interactive incremental movements. Our module can be conveniently embedded into existing multi-agent prediction methods to extend original motion distribution decoders. Extensive experiments on nuScenes and Argoverse 2 datasets demonstrate that IPCC-TP improves the performance of baselines by a large margin.
****
Improving Robust Generalization by Direct PAC-Bayesian Bound Minimization
Zifan Wang, Nan Ding, Tomer Levinboim, Xi Chen, Radu Soricut; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16458-16468
Recent research in robust optimization has shown an overfitting-like phenomenon in which models trained against adversarial attacks exhibit higher robustness on the training set compared to the test set. Although previous work provided theoretical explanations for this phenomenon using a robust PAC-Bayesian bound over the adversarial test error, related algorithmic derivations are at best only loosely connected to this bound, which implies that there is still a gap between their empirical success and our understanding of adversarial robustness theory. To close this gap, in this paper we consider a different form of the robust PAC-Bayesian bound and directly minimize it with respect to the model posterior. The derivation of the optimal solution connects PAC-Bayesian learning to the geometry of the robust loss surface through a Trace of Hessian (TrH) regularizer that measures the surface flatness. In practice, we restrict the TrH regularizer to the top layer only, which results in an analytical solution to the bound whose computational cost does not depend on the network depth. Finally, we evaluate our TrH regularization approach over CIFAR-10/100 and ImageNet using Vision Transforme

rs (ViT) and compare against baseline adversarial robustness algorithms. Experimental results show that TrH regularization leads to improved ViT robustness that either matches or surpasses previous state-of-the-art approaches while at the same time requires less memory and computational cost.
****
MobileOne: An Improved One Millisecond Mobile Backbone
Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, Anurag Ranjan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7907-7917
Efficient neural network backbones for mobile devices are often optimized for metrics such as FLOPs or parameter count. However, these metrics may not correlate well with latency of the network when deployed on a mobile device. Therefore, we perform extensive analysis of different metrics by deploying several mobile-friendly networks on a mobile device. We identify and analyze architectural and optimization bottlenecks in recent efficient neural networks and provide ways to mitigate these bottlenecks. To this end, we design an efficient backbone MobileOne, with variants achieving an inference time under 1 ms on an iPhone12 with 75.9% top-1 accuracy on ImageNet. We show that MobileOne achieves state-of-the-art performance within the efficient architectures while being many times faster on mobile. Our best model obtains similar performance on ImageNet as MobileFormer while being 38x faster. Our model obtains 2.3% better top-1 accuracy on ImageNet than EfficientNet at similar latency. Furthermore, we show that our model generalizes to multiple tasks -- image classification, object detection, and semantic segmentation with significant improvements in latency and accuracy as compared to existing efficient architectures when deployed on a mobile device.
****
A Data-Based Perspective on Transfer Learning
Saachi Jain, Hadi Salman, Alaa Khaddaj, Eric Wong, Sung Min Park, Aleksander M■dry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3613-3622
It is commonly believed that more pre-training data leads to better transfer learning performance. However, recent evidence suggests that removing data from the source dataset can actually help too. In this work, we present a framework for probing the impact of the source dataset's composition on transfer learning performance. Our framework facilitates new capabilities such as identifying transfer learning brittleness and detecting pathologies such as data-leakage and the presence of misleading examples in the source dataset. In particular, we demonstrate that removing detrimental datapoints identified by our framework improves transfer performance from ImageNet on a variety of transfer tasks.
****
AssemblyHands: Towards Egocentric Activity Understanding via 3D Hand Pose Estimation
Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, Cem Keskin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12999-13008
We present AssemblyHands, a large-scale benchmark dataset with accurate 3D hand pose annotations, to facilitate the study of egocentric activities with challenging hand-object interactions. The dataset includes synchronized egocentric and exocentric images sampled from the recent Assembly101 dataset, in which participants assemble and disassemble take-apart toys. To obtain high-quality 3D hand pose annotations for the egocentric images, we develop an efficient pipeline, where we use an initial set of manual annotations to train a model to automatically annotate a much larger dataset. Our annotation model uses multi-view feature fusion and an iterative refinement scheme, and achieves an average keypoint error of 4.20 mm, which is 85 % lower than the error of the original annotations in Assembly101. AssemblyHands provides 3.0M annotated images, including 490K egocentric images, making it the largest existing benchmark dataset for egocentric 3D hand pose estimation. Using this data, we develop a strong single-view baseline of 3D hand pose estimation from egocentric images. Furthermore, we design a novel action classification task to evaluate predicted 3D hand poses. Our study shows th

at having higher-quality hand poses directly improves the ability to recognize actions.
****
Scene-Aware Egocentric 3D Human Pose Estimation
Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, Christian Theobalt; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13031-13040
Egocentric 3D human pose estimation with a single head-mounted fisheye camera has recently attracted attention due to its numerous applications in virtual and augmented reality. Existing methods still struggle in challenging poses where the human body is highly occluded or is closely interacting with the scene. To address this issue, we propose a scene-aware egocentric pose estimation method that guides the prediction of the egocentric pose with scene constraints. To this end, we propose an egocentric depth estimation network to predict the scene depth map from a wide-view egocentric fisheye camera while mitigating the occlusion of the human body with a depth-inpainting network. Next, we propose a scene-aware pose estimation network that projects the 2D image features and estimated depth map of the scene into a voxel space and regresses the 3D pose with a V2V network. The voxel-based feature representation provides the direct geometric connection between 2D image features and scene geometry, and further facilitates the V2V network to constrain the predicted pose based on the estimated scene geometry. To enable the training of the aforementioned networks, we also generated a synthetic dataset, called EgoGTA, and an in-the-wild dataset based on EgoPW, called EgoPW-Scene. The experimental results of our new evaluation sequences show that the predicted 3D egocentric poses are accurate and physically plausible in terms of human-scene interaction, demonstrating that our method outperforms the state-of-the-art methods both quantitatively and qualitatively.
****
Learning Geometry-Aware Representations by Sketching
Hyundo Lee, Inwoo Hwang, Hyunsung Go, Won-Seok Choi, Kibeom Kim, Byoung-Tak Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23315-23326
Understanding geometric concepts, such as distance and shape, is essential for understanding the real world and also for many vision tasks. To incorporate such information into a visual representation of a scene, we propose learning to represent the scene by sketching, inspired by human behavior. Our method, coined Learning by Sketching (LBS), learns to convert an image into a set of colored strokes that explicitly incorporate the geometric information of the scene in a single inference step without requiring a sketch dataset. A sketch is then generated from the strokes where CLIP-based perceptual loss maintains a semantic similarity between the sketch and the image. We show theoretically that sketching is equivariant with respect to arbitrary affine transformations and thus provably preserves geometric information. Experimental results show that LBS substantially improves the performance of object attribute classification on the unlabeled CLEVR dataset, domain transfer between CLEVR and STL-10 datasets, and for diverse downstream tasks, confirming that LBS provides rich geometric information.
****
SVFormer: Semi-Supervised Video Transformer for Action Recognition
Zhen Xing, Qi Dai, Han Hu, Jingjing Chen, Zuxuan Wu, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18816-18826
Semi-supervised action recognition is a challenging but critical task due to the high cost of video annotations. Existing approaches mainly use convolutional neural networks, yet current revolutionary vision transformer models have been less explored. In this paper, we investigate the use of transformer models under the SSL setting for action recognition. To this end, we introduce SVFormer, which adopts a steady pseudo-labeling framework (ie, EMA-Teacher) to cope with unlabeled video samples. While a wide range of data augmentations have been shown effective for semi-supervised image classification, they generally produce limited results for video recognition. We therefore introduce a novel augmentation strateg

y, Tube TokenMix, tailored for video data where video clips are mixed via a mask with consistent masked tokens over the temporal axis. In addition, we propose a temporal warping augmentation to cover the complex temporal variation in videos, which stretches selected frames to various temporal durations in the clip. Extensive experiments on three datasets Kinetics-400, UCF-101, and HMDB-51 verify the advantage of SVFormer. In particular, SVFormer outperforms the state-of-the-art by 31.5% with fewer training epochs under the 1% labeling rate of Kinetics-400. Our method can hopefully serve as a strong benchmark and encourage future search on semi-supervised action recognition with Transformer networks.
****
X-Avatar: Expressive Human Avatars
Kaiyue Shen, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Julien Valentin, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16911-16921
We present X-Avatar, a novel avatar model that captures the full expressiveness of digital humans to bring about life-like experiences in telepresence, AR/VR and beyond. Our method models bodies, hands, facial expressions and appearance in a holistic fashion and can be learned from either full 3D scans or RGB-D data. To achieve this, we propose a part-aware learned forward skinning module that can be driven by the parameter space of SMPL-X, allowing for expressive animation of X-Avatars. To efficiently learn the neural shape and deformation fields, we propose novel part-aware sampling and initialization strategies. This leads to higher fidelity results, especially for smaller body parts while maintaining efficient training despite increased number of articulated bones. To capture the appearance of the avatar with high-frequency details, we extend the geometry and deformation fields with a texture network that is conditioned on pose, facial expression, geometry and the normals of the deformed surface. We show experimentally that our method outperforms strong baselines both quantitatively and qualitatively on the animation task. To facilitate future research on expressive avatars we contribute a new dataset, called X-Humans, containing 233 sequences of high-quality textured scans from 20 participants, totalling 35,500 data frames.
****
AccelIR: Task-Aware Image Compression for Accelerating Neural Restoration
Juncheol Ye, Hyunho Yeo, Jinwoo Park, Dongsu Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18216-18226
Recently, deep neural networks have been successfully applied for image restoration (IR) (e.g., super-resolution, de-noising, de-blurring). Despite their promising performance, running IR networks requires heavy computation. A large body of work has been devoted to addressing this issue by designing novel neural networks or pruning their parameters. However, the common limitation is that while images are saved in a compressed format before being enhanced by IR, prior work does not consider the impact of compression on the IR quality. In this paper, we present AccelIR, a framework that optimizes image compression considering the end-to-end pipeline of IR tasks. AccelIR encodes an image through IR-aware compression that optimizes compression levels across image blocks within an image according to the impact on the IR quality. Then, it runs a lightweight IR network on the compressed image, effectively reducing IR computation, while maintaining the same IR quality and image size. Our extensive evaluation using seven IR networks shows that AccelIR can reduce the computing overhead of super-resolution, de-nosing, and de-blurring by 49%, 29%, and 32% on average, respectively
****
BEV-Guided Multi-Modality Fusion for Driving Perception
Yunze Man, Liang-Yan Gui, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21960-21969
Integrating multiple sensors and addressing diverse tasks in an end-to-end algorithm are challenging yet critical topics for autonomous driving. To this end, we introduce BEVGuide, a novel Bird's Eye-View (BEV) representation learning framework, representing the first attempt to unify a wide range of sensors under direct BEV guidance in an end-to-end fashion. Our architecture accepts input from a

diverse sensor pool, including but not limited to Camera, Lidar and Radar sensors, and extracts BEV feature embeddings using a versatile and general transformer backbone. We design a BEV-guided multi-sensor attention block to take queries from BEV embeddings and learn the BEV representation from sensor-specific features. BEVGuide is efficient due to its lightweight backbone design and highly flexible as it supports almost any input sensor configurations. Extensive experiments demonstrate that our framework achieves exceptional performance in BEV perception tasks with a diverse sensor set. Project page is at https://yunzeman.github.io/BEVGuide.
****
Meta-Explore: Exploratory Hierarchical Vision-and-Language Navigation Using Scene Object Spectrum Grounding
Minyoung Hwang, Jaeyeon Jeong, Minsoo Kim, Yoonseon Oh, Songhwai Oh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6683-6693
The main challenge in vision-and-language navigation (VLN) is how to understand natural-language instructions in an unseen environment. The main limitation of conventional VLN algorithms is that if an action is mistaken, the agent fails to follow the instructions or explores unnecessary regions, leading the agent to an irrecoverable path. To tackle this problem, we propose Meta-Explore, a hierarchical navigation method deploying an exploitation policy to correct misled recent actions. We show that an exploitation policy, which moves the agent toward a well-chosen local goal among unvisited but observable states, outperforms a method which moves the agent to a previously visited state. We also highlight the demand for imagining regretful explorations with semantically meaningful clues. The key to our approach is understanding the object placements around the agent in spectral-domain. Specifically, we present a novel visual representation, called scene object spectrum (SOS), which performs category-wise 2D Fourier transform of detected objects. Combining exploitation policy and SOS features, the agent can correct its path by choosing a promising local goal. We evaluate our method in three VLN benchmarks: R2R, SOON, and REVERIE. Meta-Explore outperforms other baselines and shows significant generalization performance. In addition, local goal search using the proposed spectral-domain SOS features significantly improves the success rate by 17.1% and SPL by 20.6% for the SOON benchmark.
****
Proximal Splitting Adversarial Attack for Semantic Segmentation
Jérôme Rony, Jean-Christophe Pesquet, Ismail Ben Ayed; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20524-20533
Classification has been the focal point of research on adversarial attacks, but only a few works investigate methods suited to denser prediction tasks, such as semantic segmentation. The methods proposed in these works do not accurately solve the adversarial segmentation problem and, therefore, overestimate the size of the perturbations required to fool models. Here, we propose a white-box attack for these models based on a proximal splitting to produce adversarial perturbations with much smaller l_infinity norms. Our attack can handle large numbers of constraints within a nonconvex minimization framework via an Augmented Lagrangian approach, coupled with adaptive constraint scaling and masking strategies. We demonstrate that our attack significantly outperforms previously proposed ones, as well as classification attacks that we adapted for segmentation, providing a first comprehensive benchmark for this dense task.
****
Improved Test-Time Adaptation for Domain Generalization
Liang Chen, Yong Zhang, Yibing Song, Ying Shan, Lingqiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24172-24182
The main challenge in domain generalization (DG) is to handle the distribution shift problem that lies between the training and test data. Recent studies suggest that test-time training (TTT), which adapts the learned model with test data, might be a promising solution to the problem. Generally, a TTT strategy hinges i

ts performance on two main factors: selecting an appropriate auxiliary TTT task for updating and identifying reliable parameters to update during the test phase. Both previous arts and our experiments indicate that TTT may not improve but be detrimental to the learned model if those two factors are not properly considered. This work addresses those two factors by proposing an Improved Test-Time Adaptation (ITTA) method. First, instead of heuristically defining an auxiliary objective, we propose a learnable consistency loss for the TTT task, which contains learnable parameters that can be adjusted toward better alignment between our TTT task and the main prediction task. Second, we introduce additional adaptive parameters for the trained model, and we suggest only updating the adaptive parameters during the test phase. Through extensive experiments, we show that the proposed two strategies are beneficial for the learned model (see Figure 1), and ITTA could achieve superior performance to the current state-of-the-arts on several DG benchmarks.
****
Recovering 3D Hand Mesh Sequence From a Single Blurry Image: A New Dataset and Temporal Unfolding
Yeonguk Oh, JoonKyu Park, Jaeha Kim, Gyeongsik Moon, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 554-563
Hands, one of the most dynamic parts of our body, suffer from blur due to their active movements. However, previous 3D hand mesh recovery methods have mainly focused on sharp hand images rather than considering blur due to the absence of datasets providing blurry hand images. We first present a novel dataset BlurHand, which contains blurry hand images with 3D groundtruths. The BlurHand is constructed by synthesizing motion blur from sequential sharp hand images, imitating realistic and natural motion blurs. In addition to the new dataset, we propose BlurHandNet, a baseline network for accurate 3D hand mesh recovery from a blurry hand image. Our BlurHandNet unfolds a blurry input image to a 3D hand mesh sequence to utilize temporal information in the blurry input image, while previous works output a static single hand mesh. We demonstrate the usefulness of BlurHand for the 3D hand mesh recovery from blurry images in our experiments. The proposed BlurHandNet produces much more robust results on blurry images while generalizing well to in-the-wild images. The training codes and BlurHand dataset are available at https://github.com/JaehaKim97/BlurHand_RELEASE.
****
NaQ: Leveraging Narrations As Queries To Supervise Episodic Memory
Santhosh Kumar Ramakrishnan, Ziad Al-Halah, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6694-6703
Searching long egocentric videos with natural language queries (NLQ) has compelling applications in augmented reality and robotics, where a fluid index into everything that a person (agent) has seen before could augment human memory and surface relevant information on demand. However, the structured nature of the learning problem (free-form text query inputs, localized video temporal window outputs) and its needle-in-a-haystack nature makes it both technically challenging and expensive to supervise. We introduce Narrations-as-Queries (NaQ), a data augmentation strategy that transforms standard video-text narrations into training data for a video query localization model. Validating our idea on the Ego4D benchmark, we find it has tremendous impact in practice. NaQ improves multiple top models by substantial margins (even doubling their accuracy), and yields the very best results to date on the Ego4D NLQ challenge, soundly outperforming all challenge winners in the CVPR and ECCV 2022 competitions and topping the current public leaderboard. Beyond achieving the state-of-the-art for NLQ, we also demonstrate unique properties of our approach such as the ability to perform zero-shot and few-shot NLQ, and improved performance on queries about long-tail object categories. Code and models: http://vision.cs.utexas.edu/projects/naq.
****
Correspondence Transformers With Asymmetric Feature Learning and Matching Flow Super-Resolution

Yixuan Sun, Dongyang Zhao, Zhangyue Yin, Yiwen Huang, Tao Gui, Wenqiang Zhang, Weifeng Ge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17787-17796

This paper solves the problem of learning dense visual correspondences between different object instances of the same category with only sparse annotations. We decompose this pixel-level semantic matching problem into two easier ones: (i) First, local feature descriptors of source and target images need to be mapped into shared semantic spaces to get coarse matching flows. (ii) Second, matching flows in low resolution should be refined to generate accurate point-to-point matching results. We propose asymmetric feature learning and matching flow super-resolution based on vision transformers to solve the above problems. The asymmetric feature learning module exploits a biased cross-attention mechanism to encode token features of source images with their target counterparts. Then matching flow in low resolutions is enhanced by a super-resolution network to get accurate correspondences. Our pipeline is built upon vision transformers and can be trained in an end-to-end manner. Extensive experimental results on several popular benchmarks, such as PF-PASCAL, PF-WILLOW, and SPair-71K, demonstrate that the proposed method can catch subtle semantic differences in pixels efficiently. Code is available on https://github.com/YXSUNMADMAX/ACTR.

****

Adjustment and Alignment for Unbiased Open Set Domain Adaptation

Wuyang Li, Jie Liu, Bo Han, Yixuan Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24110-24119

Open Set Domain Adaptation (OSDA) transfers the model from a label-rich domain to a label-free one containing novel-class samples. Existing OSDA works overlook abundant novel-class semantics hidden in the source domain, leading to a biased model learning and transfer. Although the causality has been studied to remove the semantic-level bias, the non-available novel-class samples result in the failure of existing causal solutions in OSDA. To break through this barrier, we propose a novel causality-driven solution with the unexplored front-door adjustment theory, and then implement it with a theoretically grounded framework, coined AdjustmeNt aNd Alignment (ANNA), to achieve an unbiased OSDA. In a nutshell, ANNA consists of Front-Door Adjustment (FDA) to correct the biased learning in the source domain and Decoupled Causal Alignment (DCA) to transfer the model unbiasedly. On the one hand, FDA delves into fine-grained visual blocks to discover novel-class regions hidden in the base-class image. Then, it corrects the biased model optimization by implementing causal debiasing. On the other hand, DCA disentangles the base-class and novel-class regions with orthogonal masks, and then adapts the decoupled distribution for an unbiased model transfer. Extensive experiments show that ANNA achieves state-of-the-art results. The code is available at https://github.com/CityU-AIM-Group/Anna.

****

FedSeg: Class-Heterogeneous Federated Learning for Semantic Segmentation

Jiaxu Miao, Zongxin Yang, Leilei Fan, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8042-8052

Federated Learning (FL) is a distributed learning paradigm that collaboratively learns a global model across multiple clients with data privacy-preserving. Although many FL algorithms have been proposed for classification tasks, few works focus on more challenging semantic seg-mentation tasks, especially in the class-heterogeneous FL situation. Compared with classification, the issues from heterogeneous FL for semantic segmentation are more severe: (1) Due to the non-IID distribution, different clients may contain inconsistent foreground-background classes, resulting in divergent local updates. (2) Class-heterogeneity for complex dense prediction tasks makes the local optimum of clients farther from the global optimum. In this work, we propose FedSeg, a basic federated learning approach for class-heterogeneous semantic segmentation. We first propose a simple but strong modified cross-entropy loss to correct the local optimization and address the foreground-background inconsistency problem. Based on it, we introduce pixel-level contrastive learning to enforce local pixel embeddings belonging to the global semantic space. Extensive experiments on four semantic segmentation benchmarks

(Cityscapes, CamVID, PascalVOC and ADE20k) demonstrate the effectiveness of our FedSeg. We hope this work will attract more attention from the FL community to the challenging semantic segmentation federated learning.
****
NeuralField-LDM: Scene Generation With Hierarchical Latent Diffusion Models
Seung Wook Kim, Bradley Brown, Kangxue Yin, Karsten Kreis, Katja Schwarz, Daiqing Li, Robin Rombach, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8496-8506
Automatically generating high-quality real world 3D scenes is of enormous interest for applications such as virtual reality and robotics simulation. Towards this goal, we introduce NeuralField-LDM, a generative model capable of synthesizing complex 3D environments. We leverage Latent Diffusion Models that have been successfully utilized for efficient high-quality 2D content creation. We first train a scene auto-encoder to express a set of image and pose pairs as a neural field, represented as density and feature voxel grids that can be projected to produce novel views of the scene. To further compress this representation, we train a latent-autoencoder that maps the voxel grids to a set of latent representations. A hierarchical diffusion model is then fit to the latents to complete the scene generation pipeline. We achieve a substantial improvement over existing state-of-the-art scene generation models. Additionally, we show how NeuralField-LDM can be used for a variety of 3D content creation applications, including conditional scene generation, scene inpainting and scene style manipulation.
****
DPF: Learning Dense Prediction Fields With Weak Supervision
Xiaoxue Chen, Yuhang Zheng, Yupeng Zheng, Qiang Zhou, Hao Zhao, Guyue Zhou, Ya-Qin Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15347-15357
Nowadays, many visual scene understanding problems are addressed by dense prediction networks. But pixel-wise dense annotations are very expensive (e.g., for scene parsing) or impossible (e.g., for intrinsic image decomposition), motivating us to leverage cheap point-level weak supervision. However, existing pointly-supervised methods still use the same architecture designed for full supervision. In stark contrast to them, we propose a new paradigm that makes predictions for point coordinate queries, as inspired by the recent success of implicit representations, like distance or radiance fields. As such, the method is named as dense prediction fields (DPFs). DPFs generate expressive intermediate features for continuous sub-pixel locations, thus allowing outputs of an arbitrary resolution. DPFs are naturally compatible with point-level supervision. We showcase the effectiveness of DPFs using two substantially different tasks: high-level semantic parsing and low-level intrinsic image decomposition. In these two cases, supervision comes in the form of single-point semantic category and two-point relative reflectance, respectively. As benchmarked by three large-scale public datasets PascalContext, ADE20k and IIW, DPFs set new state-of-the-art performance on all of them with significant margins. Code can be accessed at https://github.com/cxx226/DPF.
****
Fast Monocular Scene Reconstruction With Global-Sparse Local-Dense Grids
Wei Dong, Christopher Choy, Charles Loop, Or Litany, Yuke Zhu, Anima Anandkumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4263-4272
Indoor scene reconstruction from monocular images has long been sought after by augmented reality and robotics developers. Recent advances in neural field representations and monocular priors have led to remarkable results in scene-level surface reconstructions. The reliance on Multilayer Perceptrons (MLP), however, significantly limits speed in training and rendering. In this work, we propose to directly use signed distance function (SDF) in sparse voxel block grids for fast and accurate scene reconstruction without MLPs. Our globally sparse and locally dense data structure exploits surfaces' spatial sparsity, enables cache-friendly queries, and allows direct extensions to multi-modal data such as color and se

mantic labels. To apply this representation to monocular scene reconstruction, w
e develop a scale calibration algorithm for fast geometric initialization from m
onocular depth priors. We apply differentiable volume rendering from this initia
lization to refine details with fast convergence. We also introduce efficient hi
gh-dimensional Continuous Random Fields (CRFs) to further exploit the semantic-g
eometry consistency between scene objects. Experiments show that our approach is
 10x faster in training and 100x faster in rendering while achieving comparable
accuracy to state-of-the-art neural implicit methods.
****
Thermal Spread Functions (TSF): Physics-Guided Material Classification
Aniket Dashpute, Vishwanath Saragadam, Emma Alexander, Florian Willomitzer, Agge
los Katsaggelos, Ashok Veeraraghavan, Oliver Cossairt; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1641-
1650
Robust and non-destructive material classification is a challenging but crucial
first-step in numerous vision applications. We propose a physics-guided material
 classification framework that relies on thermal properties of the object. Our k
ey observation is that the rate of heating and cooling of an object depends on t
he unique intrinsic properties of the material, namely the emissivity and diffus
ivity. We leverage this observation by gently heating the objects in the scene w
ith a low-power laser for a fixed duration and then turning it off, while a ther
mal camera captures measurements during the heating and cooling process. We then
 take this spatial and temporal "thermal spread function" (TSF) to solve an inve
rse heat equation using the finite-differences approach, resulting in a spatiall
y varying estimate of diffusivity and emissivity. These tuples are then used to
train a classifier that produces a fine-grained material label at each spatial p
ixel. Our approach is extremely simple requiring only a small light source (low
power laser) and a thermal camera, and produces robust classification results wi
th 86% accuracy over 16 classes
****
ESLAM: Efficient Dense SLAM System Based on Hybrid Representation of Signed Dist
ance Fields
Mohammad Mahdi Johari, Camilla Carta, François Fleuret; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1740
8-17419
We present ESLAM, an efficient implicit neural representation method for Simulta
neous Localization and Mapping (SLAM). ESLAM reads RGB-D frames with unknown cam
era poses in a sequential manner and incrementally reconstructs the scene repres
entation while estimating the current camera position in the scene. We incorpora
te the latest advances in Neural Radiance Fields (NeRF) into a SLAM system, resu
lting in an efficient and accurate dense visual SLAM method. Our scene represent
ation consists of multi-scale axis-aligned perpendicular feature planes and shal
low decoders that, for each point in the continuous space, decode the interpolat
ed features into Truncated Signed Distance Field (TSDF) and RGB values. Our exte
nsive experiments on three standard datasets, Replica, ScanNet, and TUM RGB-D sh
ow that ESLAM improves the accuracy of 3D reconstruction and camera localization
 of state-of-the-art dense visual SLAM methods by more than 50%, while it runs u
p to 10 times faster and does not require any pre-training. Project page: https:
//www.idiap.ch/paper/eslam
****
CNVid-3.5M: Build, Filter, and Pre-Train the Large-Scale Public Chinese Video-Te
xt Dataset
Tian Gan, Qing Wang, Xingning Dong, Xiangyuan Ren, Liqiang Nie, Qingpei Guo; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 14815-14824
Owing to well-designed large-scale video-text datasets, recent years have witnes
sed tremendous progress in video-text pre-training. However, existing large-scal
e video-text datasets are mostly English-only. Though there are certain methods
studying the Chinese video-text pre-training, they pre-train their models on pri
vate datasets whose videos and text are unavailable. This lack of large-scale pu

blic datasets and benchmarks in Chinese hampers the research and downstream applications of Chinese video-text pre-training. Towards this end, we release and benchmark CNVid-3.5M, a large-scale public cross-modal dataset containing over 3.5M Chinese video-text pairs. We summarize our contributions by three verbs, i.e., "Build", "Filter", and "Pre-train": 1) To build a public Chinese video-text dataset, we collect over 4.5M videos from the Chinese websites. 2) To improve the data quality, we propose a novel method to filter out 1M weakly-paired videos, resulting in the CNVid-3.5M dataset. And 3) we benchmark CNVid-3.5M with three mainstream pixel-level pre-training architectures. At last, we propose the Hard Sample Curriculum Learning strategy to promote the pre-training performance. To the best of our knowledge, CNVid-3.5M is the largest public video-text dataset in Chinese, and we provide the first pixel-level benchmarks for Chinese video-text pre-training. The dataset, codebase, and pre-trained models are available at https://github.com/CNVid/CNVid-3.5M.
****
Unsupervised Space-Time Network for Temporally-Consistent Segmentation of Multiple Motions
Etienne Meunier, Patrick Bouthemy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22139-22148
Motion segmentation is one of the main tasks in computer vision and is relevant for many applications. The optical flow (OF) is the input generally used to segment every frame of a video sequence into regions of coherent motion. Temporal consistency is a key feature of motion segmentation, but it is often neglected. In this paper, we propose an original unsupervised spatio-temporal framework for motion segmentation from optical flow that fully investigates the temporal dimension of the problem. More specifically, we have defined a 3D network for multiple motion segmentation that takes as input a sub-volume of successive optical flows and delivers accordingly a sub-volume of coherent segmentation maps. Our network is trained in a fully unsupervised way, and the loss function combines a flow reconstruction term involving spatio-temporal parametric motion models, and a regularization term enforcing temporal consistency on the masks. We have specified an easy temporal linkage of the predicted segments. Besides, we have proposed a flexible and efficient way of coding U-nets. We report experiments on several VOS benchmarks with convincing quantitative results, while not using appearance and not training with any ground-truth data. We also highlight through visual results the distinctive contribution of the short- and long-term temporal consistency brought by our OF segmentation method.
****
Unsupervised 3D Point Cloud Representation Learning by Triangle Constrained Contrast for Autonomous Driving
Bo Pang, Hongchi Xia, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5229-5239
Due to the difficulty of annotating the 3D LiDAR data of autonomous driving, an efficient unsupervised 3D representation learning method is important. In this paper, we design the Triangle Constrained Contrast (TriCC) framework tailored for autonomous driving scenes which learns 3D unsupervised representations through both the multimodal information and dynamic of temporal sequences. We treat one camera image and two LiDAR point clouds with different timestamps as a triplet. And our key design is the consistent constraint that automatically finds matching relationships among the triplet through "self-cycle" and learns representations from it. With the matching relations across the temporal dimension and modalities, we can further conduct a triplet contrast to improve learning efficiency. To the best of our knowledge, TriCC is the first framework that unifies both the temporal and multimodal semantics, which means it utilizes almost all the information in autonomous driving scenes. And compared with previous contrastive methods, it can automatically dig out contrasting pairs with higher difficulty, instead of relying on handcrafted ones. Extensive experiments are conducted with Minkowski-UNet and VoxelNet on several semantic segmentation and 3D detection datasets. Results show that TriCC learns effective representations with much fewer training iterations and improves the SOTA results greatly on all the downstream tas

ks. Code and models can be found at https://bopang1996.github.io/.
****
iDisc: Internal Discretization for Monocular Depth Estimation
Luigi Piccinelli, Christos Sakaridis, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21477-21487
Monocular depth estimation is fundamental for 3D scene understanding and downstream applications. However, even under the supervised setup, it is still challenging and ill-posed due to the lack of geometric constraints. We observe that although a scene can consist of millions of pixels, there are much fewer high-level patterns. We propose iDisc to learn those patterns with internal discretized representations. The method implicitly partitions the scene into a set of high-level concepts. In particular, our new module, Internal Discretization (ID), implements a continuous-discrete-continuous bottleneck to learn those concepts without supervision. In contrast to state-of-the-art methods, the proposed model does not enforce any explicit constraints or priors on the depth output. The whole network with the ID module can be trained in an end-to-end fashion thanks to the bottleneck module based on attention. Our method sets the new state of the art with significant improvements on NYU-Depth v2 and KITTI, outperforming all published methods on the official KITTI benchmark. iDisc can also achieve state-of-the-art results on surface normal estimation. Further, we explore the model generalization capability via zero-shot testing. From there, we observe the compelling need to promote diversification in the outdoor scenario and we introduce splits of two autonomous driving datasets, DDAD and Argoverse. Code is available at http://vis.xyz/pub/idisc/.
****
Balancing Logit Variation for Long-Tailed Semantic Segmentation
Yuchao Wang, Jingjing Fei, Haochen Wang, Wei Li, Tianpeng Bao, Liwei Wu, Rui Zhao, Yujun Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19561-19573
Semantic segmentation usually suffers from a long tail data distribution. Due to the imbalanced number of samples across categories, the features of those tail classes may get squeezed into a narrow area in the feature space. Towards a balanced feature distribution, we introduce category-wise variation into the network predictions in the training phase such that an instance is no longer projected to a feature point, but a small region instead. Such a perturbation is highly dependent on the category scale, which appears as assigning smaller variation to head classes and larger variation to tail classes. In this way, we manage to close the gap between the feature areas of different categories, resulting in a more balanced representation. It is noteworthy that the introduced variation is discarded at the inference stage to facilitate a confident prediction. Although with an embarrassingly simple implementation, our method manifests itself in strong generalizability to various datasets and task settings. Extensive experiments suggest that our plug-in design lends itself well to a range of state-of-the-art approaches and boosts the performance on top of them.
****
Prompt-Guided Zero-Shot Anomaly Action Recognition Using Pretrained Deep Skeleton Features
Fumiaki Sato, Ryo Hachiuma, Taiki Sekii; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6471-6480
This study investigates unsupervised anomaly action recognition, which identifies video-level abnormal-human-behavior events in an unsupervised manner without abnormal samples, and simultaneously addresses three limitations in the conventional skeleton-based approaches: target domain-dependent DNN training, robustness against skeleton errors, and a lack of normal samples. We present a unified, user prompt-guided zero-shot learning framework using a target domain-independent skeleton feature extractor, which is pretrained on a large-scale action recognition dataset. Particularly, during the training phase using normal samples, the method models the distribution of skeleton features of the normal actions while freezing the weights of the DNNs and estimates the anomaly score using this distribution in the inference phase. Additionally, to increase robustness against skel

eton errors, we introduce a DNN architecture inspired by a point cloud deep lear
ning paradigm, which sparsely propagates the features between joints. Furthermor
e, to prevent the unobserved normal actions from being misidentified as abnormal
 actions, we incorporate a similarity score between the user prompt embeddings a
nd skeleton features aligned in the common space into the anomaly score, which i
ndirectly supplements normal actions. On two publicly available datasets, we con
duct experiments to test the effectiveness of the proposed method with respect t
o abovementioned limitations.
****
iQuery: Instruments As Queries for Audio-Visual Sound Separation
Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, Jianbo Shi; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2023, pp. 14675-14686
Current audio-visual separation methods share a standard architecture design whe
re an audio encoder-decoder network is fused with visual encoding features at th
e encoder bottleneck. This design confounds the learning of multi-modal feature
encoding with robust sound decoding for audio separation. To generalize to a new
 instrument, one must fine-tune the entire visual and audio network for all musi
cal instruments. We re-formulate the visual-sound separation task and propose In
struments as Queries (iQuery) with a flexible query expansion mechanism. Our app
roach ensures cross-modal consistency and cross-instrument disentanglement. We u
tilize "visually named" queries to initiate the learning of audio queries and us
e cross-modal attention to remove potential sound source interference at the est
imated waveforms. To generalize to a new instrument or event class, drawing insp
iration from the text-prompt design, we insert additional queries as audio promp
ts while freezing the attention mechanism. Experimental results on three benchma
rks demonstrate that our iQuery improves audio-visual sound source separation pe
rformance. Code is available at https://github.com/JiabenChen/iQuery.
****
Sampling Is Matter: Point-Guided 3D Human Mesh Reconstruction
Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, Wonjun Kim
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2023, pp. 12880-12889
This paper presents a simple yet powerful method for 3D human mesh reconstructio
n from a single RGB image. Most recently, the non-local interactions of the whol
e mesh vertices have been effectively estimated in the transformer while the rel
ationship between body parts also has begun to be handled via the graph model. E
ven though those approaches have shown the remarkable progress in 3D human mesh
reconstruction, it is still difficult to directly infer the relationship between
 features, which are encoded from the 2D input image, and 3D coordinates of each
 vertex. To resolve this problem, we propose to design a simple feature sampling
 scheme. The key idea is to sample features in the embedded space by following t
he guide of points, which are estimated as projection results of 3D mesh vertice
s (i.e., ground truth). This helps the model to concentrate more on vertex-relev
ant features in the 2D space, thus leading to the reconstruction of the natural
human pose. Furthermore, we apply progressive attention masking to precisely est
imate local interactions between vertices even under severe occlusions. Experime
ntal results on benchmark datasets show that the proposed method efficiently imp
roves the performance of 3D human mesh reconstruction. The code and model are pu
blicly available at: https://github.com/DCVL-3D/PointHMR_release.
****
Efficient Multimodal Fusion via Interactive Prompting
Yaowei Li, Ruijie Quan, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2604-2613
Large-scale pre-training has brought unimodal fields such as computer vision and
 natural language processing to a new era. Following this trend, the size of mul
timodal learning models constantly increases, leading to an urgent need to reduc
e the massive computational cost of fine-tuning these models for downstream task
s. In this paper, we propose an efficient and flexible multimodal fusion method,
 namely PMF, tailored for fusing unimodally pretrained transformers. Specificall

y, we first present a modular multimodal fusion framework that exhibits high flexibility and facilitates mutual interactions among different modalities. In addition, we disentangle vanilla prompts into three types in order to learn different optimizing objectives for multimodal learning. It is also worth noting that we propose to add prompt vectors only on the deep layers of the unimodal transformers, thus significantly reducing the training memory usage. Experiment results show that our proposed method achieves comparable performance to several other multimodal finetuning methods with less than 3% trainable parameters and up to 66% saving of training memory usage.
****
Look Around for Anomalies: Weakly-Supervised Anomaly Detection via Context-Motion Relational Learning
MyeongAh Cho, Minjung Kim, Sangwon Hwang, Chaewon Park, Kyungjae Lee, Sangyoun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12137-12146
Weakly-supervised Video Anomaly Detection is the task of detecting frame-level anomalies using video-level labeled training data. It is difficult to explore class representative features using minimal supervision of weak labels with a single backbone branch. Furthermore, in real-world scenarios, the boundary between normal and abnormal is ambiguous and varies depending on the situation. For example, even for the same motion of running person, the abnormality varies depending on whether the surroundings are a playground or a roadway. Therefore, our aim is to extract discriminative features by widening the relative gap between classes' features from a single branch. In the proposed Class-Activate Feature Learning (CLAV), the features are extracted as per the weights that are implicitly activated depending on the class, and the gap is then enlarged through relative distance learning. Furthermore, as the relationship between context and motion is important in order to identify the anomalies in complex and diverse scenes, we propose a Context--Motion Interrelation Module (CoMo), which models the relationship between the appearance of the surroundings and motion, rather than utilizing only temporal dependencies or motion information. The proposed method shows SOTA performance on four benchmarks including large-scale real-world datasets, and we demonstrate the importance of relational information by analyzing the qualitative results and generalization ability.
****
Depth Estimation From Indoor Panoramas With Neural Scene Representation
Wenjie Chang, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 899-908
Depth estimation from indoor panoramas is challenging due to the equirectangular distortions of panoramas and inaccurate matching. In this paper, we propose a practical framework to improve the accuracy and efficiency of depth estimation from multi-view indoor panoramic images with the Neural Radiance Field technology. Specifically, we develop two networks to implicitly learn the Signed Distance Function for depth measurements and the radiance field from panoramas. We also introduce a novel spherical position embedding scheme to achieve high accuracy. For better convergence, we propose an initialization method for the network weights based on the Manhattan World Assumption. Furthermore, we devise a geometric consistency loss, leveraging the surface normal, to further refine the depth estimation. The experimental results demonstrate that our proposed method outperforms state-of-the-art works by a large margin in both quantitative and qualitative evaluations. Our source code is available at https://github.com/WJ-Chang-42/IndoorPanoDepth.
****
Task-Specific Fine-Tuning via Variational Information Bottleneck for Weakly-Supervised Pathology Whole Slide Image Classification
Honglin Li, Chenglu Zhu, Yunlong Zhang, Yuxuan Sun, Zhongyi Shui, Wenwei Kuang, Sunyi Zheng, Lin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7454-7463
While Multiple Instance Learning (MIL) has shown promising results in digital Pathology Whole Slide Image (WSI) analysis, such a paradigm still faces performanc

e and generalization problems due to high computational costs and limited supervision of Gigapixel WSIs. To deal with the computation problem, previous methods utilize a frozen model pretrained from ImageNet to obtain representations, however, it may lose key information owing to the large domain gap and hinder the generalization ability without image-level training-time augmentation. Though Self-supervised Learning (SSL) proposes viable representation learning schemes, the downstream task-specific features via partial label tuning are not explored. To alleviate this problem, we propose an efficient WSI fine-tuning framework motivated by the Information Bottleneck theory. The theory enables the framework to find the minimal sufficient statistics of WSI, thus supporting us to fine-tune the backbone into a task-specific representation only depending on WSI-level weak labels. The WSI-MIL problem is further analyzed to theoretically deduce our fine-tuning method. We evaluate the method on five pathological WSI datasets on various WSI heads. The experimental results show significant improvements in both accuracy and generalization compared with previous works. Source code will be available at https://github.com/invoker-LL/WSI-finetuning.
****

Detecting Everything in the Open World: Towards Universal Object Detection
Zhenyu Wang, Yali Li, Xi Chen, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao, Shengjin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11433-11443
In this paper, we formally address universal object detection, which aims to detect every scene and predict every category. The dependence on human annotations, the limited visual information, and the novel categories in the open world severely restrict the universality of traditional detectors. We propose UniDetector, a universal object detector that has the ability to recognize enormous categories in the open world. The critical points for the universality of UniDetector are: 1) it leverages images of multiple sources and heterogeneous label spaces for training through the alignment of image and text spaces, which guarantees sufficient information for universal representations. 2) it generalizes to the open world easily while keeping the balance between seen and unseen classes, thanks to abundant information from both vision and language modalities. 3) it further promotes the generalization ability to novel categories through our proposed decoupling training manner and probability calibration. These contributions allow UniDetector to detect over 7k categories, the largest measurable category size so far, with only about 500 classes participating in training. Our UniDetector behaves the strong zero-shot generalization ability on large-vocabulary datasets like LVIS, ImageNetBoxes, and VisualGenome - it surpasses the traditional supervised baselines by more than 4% on average without seeing any corresponding images. On 13 public detection datasets with various scenes, UniDetector also achieves state-of-the-art performance with only a 3% amount of training data.
****

Single Image Depth Prediction Made Better: A Multivariate Gaussian Take
Ce Liu, Suryansh Kumar, Shuhang Gu, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17346-17356
Neural-network-based single image depth prediction (SIDP) is a challenging task where the goal is to predict the scene's per-pixel depth at test time. Since the problem, by definition, is ill-posed, the fundamental goal is to come up with an approach that can reliably model the scene depth from a set of training examples. In the pursuit of perfect depth estimation, most existing state-of-the-art learning techniques predict a single scalar depth value per-pixel. Yet, it is well-known that the trained model has accuracy limits and can predict imprecise depth. Therefore, an SIDP approach must be mindful of the expected depth variations in the model's prediction at test time. Accordingly, we introduce an approach that performs continuous modeling of per-pixel depth, where we can predict and reason about the per-pixel depth and its distribution. To this end, we model per-pixel scene depth using a multivariate Gaussian distribution. Moreover, contrary to the existing uncertainty modeling methods---in the same spirit, where per-pixel depth is assumed to be independent, we introduce per-pixel covariance modelin

g that encodes its depth dependency w.r.t. all the scene points. Unfortunately, per-pixel depth covariance modeling leads to a computationally expensive continu ous loss function, which we solve efficiently using the learned low-rank approxi mation of the overall covariance matrix. Notably, when tested on benchmark datas ets such as KITTI, NYU, and SUN-RGB-D, the SIDP model obtained by optimizing our loss function shows state-of-the-art results. Our method's accuracy (named MG) is among the top on the KITTI depth-prediction benchmark leaderboard.
****
NUWA-LIP: Language-Guided Image Inpainting With Defect-Free VQGAN
Minheng Ni, Xiaoming Li, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14183-14192
Language-guided image inpainting aims to fill the defective regions of an image under the guidance of text while keeping the non-defective regions unchanged. Ho wever, directly encoding the defective images is prone to have an adverse effect on the non-defective regions, giving rise to distorted structures on non-defect ive parts. To better adapt the text guidance to the inpainting task, this paper proposes NUWA-LIP, which involves defect-free VQGAN (DF-VQGAN) and a multi-persp ective sequence-to-sequence module (MP-S2S). To be specific, DF-VQGAN introduces relative estimation to carefully control the receptive spreading, as well as sy mmetrical connections to protect structure details unchanged. For harmoniously e mbedding text guidance into the locally defective regions, MP-S2S is employed by aggregating the complementary perspectives from low-level pixels, high-level to kens as well as the text description. Experiments show that our DF-VQGAN effecti vely aids the inpainting process while avoiding unexpected changes in non-defect ive regions. Results on three open-domain benchmarks demonstrate the superior pe rformance of our method against state-of-the-arts. Our code, datasets, and model will be made publicly available.
****
One-Shot Model for Mixed-Precision Quantization
Ivan Koryakovskiy, Alexandra Yakovleva, Valentin Buchnev, Temur Isaev, Gleb Odin okikh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2023, pp. 7939-7949
Neural network quantization is a popular approach for model compression. Modern hardware supports quantization in mixed-precision mode, which allows for greater compression rates but adds the challenging task of searching for the optimal bi t width. The majority of existing searchers find a single mixed-precision archit ecture. To select an architecture that is suitable in terms of performance and r esource consumption, one has to restart searching multiple times. We focus on a specific class of methods that find tensor bit width using gradient-based optimi zation. First, we theoretically derive several methods that were empirically pro posed earlier. Second, we present a novel One-Shot method that finds a diverse s et of Pareto-front architectures in O(1) time. For large models, the proposed me thod is 5 times more efficient than existing methods. We verify the method on tw o classification and super-resolution models and show above 0.93 correlation sco re between the predicted and actual model performance. The Pareto-front architec ture selection is straightforward and takes only 20 to 40 supernet evaluations, which is the new state-of-the-art result to the best of our knowledge.
****
MARLIN: Masked Autoencoder for Facial Video Representation LearnINg
Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezat ofighi, Reza Haffari, Munawar Hayat; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2023, pp. 1493-1504
This paper proposes a self-supervised approach to learn universal facial represe ntations from videos, that can transfer across a variety of facial analysis task s such as Facial Attribute Recognition (FAR), Facial Expression Recognition (FER), DeepFake Detection (DFD), and Lip Synchronization (LS). Our proposed framewor k, named MARLIN, is a facial video masked autoencoder, that learns highly robust and generic facial embeddings from abundantly available non-annotated web crawl ed facial videos. As a challenging auxiliary task, MARLIN reconstructs the spati o-temporal details of the face from the densely masked facial regions which main

ly include eyes, nose, mouth, lips, and skin to capture local and global aspects that in turn help in encoding generic and transferable features. Through a variety of experiments on diverse downstream tasks, we demonstrate MARLIN to be an excellent facial video encoder as well as feature extractor, that performs consistently well across a variety of downstream tasks including FAR (1.13% gain over supervised benchmark), FER (2.64% gain over unsupervised benchmark), DFD (1.86% gain over unsupervised benchmark), LS (29.36% gain for Frechet Inception Distance), and even in low data regime. Our code and models are available at https://github.com/ControlNet/MARLIN.
****
Language Adaptive Weight Generation for Multi-Task Visual Grounding
Wei Su, Peihan Miao, Huanzhang Dou, Gaoang Wang, Liang Qiao, Zheyang Li, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10857-10866
Although the impressive performance in visual grounding, the prevailing approaches usually exploit the visual backbone in a passive way, i.e., the visual backbone extracts features with fixed weights without expression-related hints. The passive perception may lead to mismatches (e.g., redundant and missing), limiting further performance improvement. Ideally, the visual backbone should actively extract visual features since the expressions already provide the blueprint of desired visual features. The active perception can take expressions as priors to extract relevant visual features, which can effectively alleviate the mismatches. Inspired by this, we propose an active perception Visual Grounding framework based on Language Adaptive Weights, called VG-LAW. The visual backbone serves as an expression-specific feature extractor through dynamic weights generated for various expressions. Benefiting from the specific and relevant visual features extracted from the language-aware visual backbone, VG-LAW does not require additional modules for cross-modal interaction. Along with a neat multi-task head, VG-LAW can be competent in referring expression comprehension and segmentation jointly. Extensive experiments on four representative datasets, i.e., RefCOCO, RefCOCO+, RefCOCOg, and ReferItGame, validate the effectiveness of the proposed framework and demonstrate state-of-the-art performance.
****
Continuous Intermediate Token Learning With Implicit Motion Manifold for Keyframe Based Motion Interpolation
Clinton A. Mo, Kun Hu, Chengjiang Long, Zhiyong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13894-13903
Deriving sophisticated 3D motions from sparse keyframes is a particularly challenging problem, due to continuity and exceptionally skeletal precision. The action features are often derivable accurately from the full series of keyframes, and thus, leveraging the global context with transformers has been a promising data-driven embedding approach. However, existing methods are often with inputs of interpolated intermediate frame for continuity using basic interpolation methods with keyframes, which result in a trivial local minimum during training. In this paper, we propose a novel framework to formulate latent motion manifolds with keyframe-based constraints, from which the continuous nature of intermediate token representations is considered. Particularly, our proposed framework consists of two stages for identifying a latent motion subspace, i.e., a keyframe encoding stage and an intermediate token generation stage, and a subsequent motion synthesis stage to extrapolate and compose motion data from manifolds. Through our extensive experiments conducted on both the LaFAN1 and CMU Mocap datasets, our proposed method demonstrates both superior interpolation accuracy and high visual similarity to ground truth motions.
****
Dynamic Coarse-To-Fine Learning for Oriented Tiny Object Detection
Chang Xu, Jian Ding, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, Gui-Song Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7318-7328
Detecting arbitrarily oriented tiny objects poses intense challenges to existing

detectors, especially for label assignment. Despite the exploration of adaptive label assignment in recent oriented object detectors, the extreme geometry shape and limited feature of oriented tiny objects still induce severe mismatch and imbalance issues. Specifically, the position prior, positive sample feature, and instance are mismatched, and the learning of extreme-shaped objects is biased and unbalanced due to little proper feature supervision. To tackle these issues, we propose a dynamic prior along with the coarse-to-fine assigner, dubbed DCFL. For one thing, we model the prior, label assignment, and object representation all in a dynamic manner to alleviate the mismatch issue. For another, we leverage the coarse prior matching and finer posterior constraint to dynamically assign labels, providing appropriate and relatively balanced supervision for diverse instances. Extensive experiments on six datasets show substantial improvements to the baseline. Notably, we obtain the state-of-the-art performance for one-stage detectors on the DOTA-v1.5, DOTA-v2.0, and DIOR-R datasets under single-scale training and testing. Codes are available at https://github.com/Chasel-Tsui/mmrotate-dcfl.
****
Controllable Mesh Generation Through Sparse Latent Point Diffusion Models
Zhaoyang Lyu, Jinyi Wang, Yuwei An, Ya Zhang, Dahua Lin, Bo Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 271-280
Mesh generation is of great value in various applications involving computer graphics and virtual content, yet designing generative models for meshes is challenging due to their irregular data structure and inconsistent topology of meshes in the same category. In this work, we design a novel sparse latent point diffusion model for mesh generation. Our key insight is to regard point clouds as an intermediate representation of meshes, and model the distribution of point clouds instead. While meshes can be generated from point clouds via techniques like Shape as Points (SAP), the challenges of directly generating meshes can be effectively avoided. To boost the efficiency and controllability of our mesh generation method, we propose to further encode point clouds to a set of sparse latent points with point-wise semantic meaningful features, where two DDPMs are trained in the space of sparse latent points to respectively model the distribution of the latent point positions and features at these latent points. We find that sampling in this latent space is faster than directly sampling dense point clouds. Moreover, the sparse latent points also enable us to explicitly control both the overall structures and local details of the generated meshes. Extensive experiments are conducted on the ShapeNet dataset, where our proposed sparse latent point diffusion model achieves superior performance in terms of generation quality and controllability when compared to existing methods.
****
Query-Centric Trajectory Prediction
Zikang Zhou, Jianping Wang, Yung-Hui Li, Yu-Kai Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17863-17873
Predicting the future trajectories of surrounding agents is essential for autonomous vehicles to operate safely. This paper presents QCNet, a modeling framework toward pushing the boundaries of trajectory prediction. First, we identify that the agent-centric modeling scheme used by existing approaches requires re-normalizing and re-encoding the input whenever the observation window slides forward, leading to redundant computations during online prediction. To overcome this limitation and achieve faster inference, we introduce a query-centric paradigm for scene encoding, which enables the reuse of past computations by learning representations independent of the global spacetime coordinate system. Sharing the invariant scene features among all target agents further allows the parallelism of multi-agent trajectory decoding. Second, even given rich encodings of the scene, existing decoding strategies struggle to capture the multimodality inherent in agents' future behavior, especially when the prediction horizon is long. To tackle this challenge, we first employ anchor-free queries to generate trajectory proposals in a recurrent fashion, which allows the model to utilize different scen

e contexts when decoding waypoints at different horizons. A refinement module then takes the trajectory proposals as anchors and leverages anchor-based queries to refine the trajectories further. By supplying adaptive and high-quality anchors to the refinement module, our query-based decoder can better deal with the multimodality in the output of trajectory prediction. Our approach ranks 1st on Argoverse 1 and Argoverse 2 motion forecasting benchmarks, outperforming all methods on all main metrics by a large margin. Meanwhile, our model can achieve streaming scene encoding and parallel multi-agent decoding thanks to the query-centric design ethos.
****
The Enemy of My Enemy Is My Friend: Exploring Inverse Adversaries for Improving Adversarial Training
Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, Xiaohua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24678-24687
Although current deep learning techniques have yielded superior performance on various computer vision tasks, yet they are still vulnerable to adversarial examples. Adversarial training and its variants have been shown to be the most effective approaches to defend against adversarial examples. A particular class of these methods regularize the difference between output probabilities for an adversarial and its corresponding natural example. However, it may have a negative impact if a natural example is misclassified. To circumvent this issue, we propose a novel adversarial training scheme that encourages the model to produce similar output probabilities for an adversarial example and its "inverse adversarial" counterpart. Particularly, the counterpart is generated by maximizing the likelihood in the neighborhood of the natural example. Extensive experiments on various vision datasets and architectures demonstrate that our training method achieves state-of-the-art robustness as well as natural accuracy among robust models. Furthermore, using a universal version of inverse adversarial examples, we improve the performance of single-step adversarial training techniques at a low computational cost.
****
Look Before You Match: Instance Understanding Matters in Video Object Segmentation
Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Chuanxin Tang, Xiyang Dai, Yucheng Zhao, Yujia Xie, Lu Yuan, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2268-2278
Exploring dense matching between the current frame and past frames for long-range context modeling, memory-based methods have demonstrated impressive results in video object segmentation (VOS) recently. Nevertheless, due to the lack of instance understanding ability, the above approaches are oftentimes brittle to large appearance variations or viewpoint changes resulted from the movement of objects and cameras. In this paper, we argue that instance understanding matters in VOS, and integrating it with memory-based matching can enjoy the synergy, which is intuitively sensible from the definition of VOS task, i.e., identifying and segmenting object instances within the video. Towards this goal, we present a two-branch network for VOS, where the query-based instance segmentation (IS) branch delves into the instance details of the current frame and the VOS branch performs spatial-temporal matching with the memory bank. We employ the well-learned object queries from IS branch to inject instance-specific information into the query key, with which the instance-augmented matching is further performed. In addition, we introduce a multi-path fusion block to effectively combine the memory readout with multi-scale features from the instance segmentation decoder, which incorporates high-resolution instance-aware features to produce final segmentation results. Our method achieves state-of-the-art performance on DAVIS 2016/2017 val (92.6% and 87.1%), DAVIS 2017 test-dev (82.8%), and YouTube-VOS 2018/2019 val (86.3% and 86.3%), outperforming alternative methods by clear margins.
****
SGLoc: Scene Geometry Encoding for Outdoor LiDAR Localization
Wen Li, Shangshu Yu, Cheng Wang, Guosheng Hu, Siqi Shen, Chenglu Wen; Proceeding

s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9286-9295
LiDAR-based absolute pose regression estimates the global pose through a deep network in an end-to-end manner, achieving impressive results in learning-based localization. However, the accuracy of existing methods still has room to improve due to the difficulty of effectively encoding the scene geometry and the unsatisfactory quality of the data. In this work, we propose a novel LiDAR localization framework, SGLoc, which decouples the pose estimation to point cloud correspondence regression and pose estimation via this correspondence. This decoupling effectively encodes the scene geometry because the decoupled correspondence regression step greatly preserves the scene geometry, leading to significant performance improvement. Apart from this decoupling, we also design a tri-scale spatial feature aggregation module and inter-geometric consistency constraint loss to effectively capture scene geometry. Moreover, we empirically find that the ground truth might be noisy due to GPS/INS measuring errors, greatly reducing the pose estimation performance. Thus, we propose a pose quality evaluation and enhancement method to measure and correct the ground truth pose. Extensive experiments on the Oxford Radar RobotCar and NCLT datasets demonstrate the effectiveness of SGLoc, which outperforms state-of-the-art regression-based localization methods by 68.5% and 67.6% on position accuracy, respectively.
****
Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary
Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, Chen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7766-7775
The practical needs of the "right to be forgotten" and poisoned data removal call for efficient machine unlearning techniques, which enable machine learning models to unlearn, or to forget a fraction of training data and its lineage. Recent studies on machine unlearning for deep neural networks (DNNs) attempt to destroy the influence of the forgetting data by scrubbing the model parameters. However, it is prohibitively expensive due to the large dimension of the parameter space. In this paper, we refocus our attention from the parameter space to the decision space of the DNN model, and propose Boundary Unlearning, a rapid yet effective way to unlearn an entire class from a trained DNN model. The key idea is to shift the decision boundary of the original DNN model to imitate the decision behavior of the model retrained from scratch. We develop two novel boundary shift methods, namely Boundary Shrink and Boundary Expanding, both of which can rapidly achieve the utility and privacy guarantees. We extensively evaluate Boundary Unlearning on CIFAR-10 and Vggface2 datasets, and the results show that Boundary Unlearning can effectively forget the forgetting class on image classification and face recognition tasks, with an expected speed-up of 17x and 19x, respectively, compared with retraining from the scratch.
****
Bridging Search Region Interaction With Template for RGB-T Tracking
Tianrui Hui, Zizheng Xun, Fengguang Peng, Junshi Huang, Xiaoming Wei, Xiaolin Wei, Jiao Dai, Jizhong Han, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13630-13639
RGB-T tracking aims to leverage the mutual enhancement and complement ability of RGB and TIR modalities for improving the tracking process in various scenarios, where cross-modal interaction is the key component. Some previous methods concatenate the RGB and TIR search region features directly to perform a coarse interaction process with redundant background noises introduced. Many other methods sample candidate boxes from search frames and conduct various fusion approaches on isolated pairs of RGB and TIR boxes, which limits the cross-modal interaction within local regions and brings about inadequate context modeling. To alleviate these limitations, we propose a novel Template-Bridged Search region Interaction (TBSI) module which exploits templates as the medium to bridge the cross-modal interaction between RGB and TIR search regions by gathering and distributing target-relevant object and environment contexts. Original templates are also update

d with enriched multimodal contexts from the template medium. Our TBSI module is inserted into a ViT backbone for joint feature extraction, search-template matching, and cross-modal interaction. Extensive experiments on three popular RGB-T tracking benchmarks demonstrate our method achieves new state-of-the-art performances. Code is available at https://github.com/RyanHTR/TBSI.
****
Indescribable Multi-Modal Spatial Evaluator
Lingke Kong, X. Sharon Qi, Qijin Shen, Jiacheng Wang, Jingyi Zhang, Yanle Hu, Qichao Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9853-9862
Multi-modal image registration spatially aligns two images with different distributions. One of its major challenges is that images acquired from different imaging machines have different imaging distributions, making it difficult to focus only on the spatial aspect of the images and ignore differences in distributions. In this study, we developed a self-supervised approach, Indescribable Multi-model Spatial Evaluator (IMSE), to address multi-modal image registration. IMSE creates an accurate multi-modal spatial evaluator to measure spatial differences between two images, and then optimizes registration by minimizing the error predicted of the evaluator. To optimize IMSE performance, we also proposed a new style enhancement method called Shuffle Remap which randomizes the image distribution into multiple segments, and then randomly disorders and remaps these segments, so that the distribution of the original image is changed. Shuffle Remap can help IMSE to predict the difference in spatial location from unseen target distributions. Our results show that IMSE outperformed the existing methods for registration using T1-T2 and CT-MRI datasets. IMSE also can be easily integrated into the traditional registration process, and can provide a convenient way to evaluate and visualize registration results. IMSE also has the potential to be used as a new paradigm for image-to-image translation. Our code is available at https://github.com/Kid-Liet/IMSE.
****
ImageBind: One Embedding Space To Bind Them All
Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15180-15190
We present ImageBind, an approach to learn a joint embedding across six different modalities - images, text, audio, depth, thermal, and IMU data. We show that all combinations of paired data are not necessary to train such a joint embedding, and only image-paired data is sufficient to bind the modalities together. ImageBind can leverage recent large scale vision-language models, and extends their zero-shot capabilities to new modalities just by using their natural pairing with images. It enables novel emergent applications 'out-of-the-box' including cross-modal retrieval, composing modalities with arithmetic, cross-modal detection and generation. The emergent capabilities improve with the strength of the image encoder and we set a new state-of-the-art on emergent zero-shot recognition tasks across modalities, outperforming specialist supervised models. Finally, we show strong few-shot recognition results outperforming prior work, and that ImageBind serves as a new way to evaluate vision models for visual and non-visual tasks.
****
Orthogonal Annotation Benefits Barely-Supervised Medical Image Segmentation
Heng Cai, Shumeng Li, Lei Qi, Qian Yu, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3302-3311
Recent trends in semi-supervised learning have significantly boosted the performance of 3D semi-supervised medical image segmentation. Compared with 2D images, 3D medical volumes involve information from different directions, e.g., transverse, sagittal, and coronal planes, so as to naturally provide complementary views. These complementary views and the intrinsic similarity among adjacent 3D slices inspire us to develop a novel annotation way and its corresponding semi-supervised model for effective segmentation. Specifically, we firstly propose the orth

ogonal annotation by only labeling two orthogonal slices in a labeled volume, wh
ich significantly relieves the burden of annotation. Then, we perform registrati
on to obtain the initial pseudo labels for sparsely labeled volumes. Subsequentl
y, by introducing unlabeled volumes, we propose a dual-network paradigm named De
nse-Sparse Co-training (DeSCO) that exploits dense pseudo labels in early stage
and sparse labels in later stage and meanwhile forces consistent output of two n
etworks. Experimental results on three benchmark datasets validated our effectiv
eness in performance and efficiency in annotation. For example, with only 10 ann
otated slices, our method reaches a Dice up to 86.93% on KiTS19 dataset.
****
Exploring Motion Ambiguity and Alignment for High-Quality Video Frame Interpolat
ion
Kun Zhou, Wenbo Li, Xiaoguang Han, Jiangbo Lu; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22169-22179
For video frame interpolation(VFI), existing deep-learning-based approaches stro
ngly rely on the ground-truth (GT) intermediate frames, which sometimes ignore t
he non-unique nature of motion judging from the given adjacent frames. As a resu
lt, these methods tend to produce averaged solutions that are not clear enough.
To alleviate this issue, we propose to relax the requirement of reconstructing a
n intermediate frame as close to the GT as possible. Towards this end, we develo
p a texture consistency loss (TCL) upon the assumption that the interpolated con
tent should maintain similar structures with their counterparts in the given fra
mes. Predictions satisfying this constraint are encouraged, though they may diff
er from the predefined GT. Without the bells and whistles, our plug-and-play TCL
 is capable of improving the performance of existing VFI frameworks consistently
. On the other hand, previous methods usually adopt the cost volume or correlati
on map to achieve more accurate image or feature warping. However, the O(N^2) (N
 refers to the pixel count) computational complexity makes it infeasible for hig
h-resolution cases. In this work, we design a simple, efficient O(N) yet powerfu
l guided cross-scale pyramid alignment(GCSPA) module, where multi-scale informat
ion is highly exploited. Extensive experiments justify the efficiency and effect
iveness of the proposed strategy.
****
Knowledge Distillation for 6D Pose Estimation by Aligning Distributions of Local
 Predictions
Shuxuan Guo, Yinlin Hu, Jose M. Alvarez, Mathieu Salzmann; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1
8633-18642
Knowledge distillation facilitates the training of a compact student network by
using a deep teacher one. While this has achieved great success in many tasks, i
t remains completely unstudied for image-based 6D object pose estimation. In thi
s work, we introduce the first knowledge distillation method driven by the 6D po
se estimation task. To this end, we observe that most modern 6D pose estimation
frameworks output local predictions, such as sparse 2D keypoints or dense repres
entations, and that the compact student network typically struggles to predict s
uch local quantities precisely. Therefore, instead of imposing prediction-to-pre
diction supervision from the teacher to the student, we propose to distill the t
eacher's distribution of local predictions into the student network, facilitatin
g its training. Our experiments on several benchmarks show that our distillation
 method yields state-of-the-art results with different compact student models an
d for both keypoint-based and dense prediction-based architectures.
****
Three Guidelines You Should Know for Universally Slimmable Self-Supervised Learn
ing
Yun-Hao Cao, Peiqin Sun, Shuchang Zhou; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15742-15751
We propose universally slimmable self-supervised learning (dubbed as US3L) to ac
hieve better accuracy-efficiency trade-offs for deploying self-supervised models
 across different devices. We observe that direct adaptation of self-supervised
learning (SSL) to universally slimmable networks misbehaves as the training proc

ess frequently collapses. We then discover that temporal consistent guidance is the key to the success of SSL for universally slimmable networks, and we propose three guidelines for the loss design to ensure this temporal consistency from a unified gradient perspective. Moreover, we propose dynamic sampling and group regularization strategies to simultaneously improve training efficiency and accuracy. Our US3L method has been empirically validated on both convolutional neural networks and vision transformers. With only once training and one copy of weights, our method outperforms various state-of-the-art methods (individually trained or not) on benchmarks including recognition, object detection and instance segmentation.
****
Adaptive Annealing for Robust Geometric Estimation
Chitturi Sidhartha, Lalit Manam, Venu Madhav Govindu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21929-21939
Geometric estimation problems in vision are often solved via minimization of statistical loss functions which account for the presence of outliers in the observations. The corresponding energy landscape often has many local minima. Many approaches attempt to avoid local minima by annealing the scale parameter of loss functions using methods such as graduated non-convexity (GNC). However, little attention has been paid to the annealing schedule, which is often carried out in a fixed manner, resulting in a poor speed-accuracy trade-off and unreliable convergence to the global minimum. In this paper, we propose a principled approach for adaptively annealing the scale for GNC by tracking the positive-definiteness (i.e. local convexity) of the Hessian of the cost function. We illustrate our approach using the classic problem of registering 3D correspondences in the presence of noise and outliers. We also develop approximations to the Hessian that significantly speeds up our method. The effectiveness of our approach is validated by comparing its performance with state-of-the-art 3D registration approaches on a number of synthetic and real datasets. Our approach is accurate and efficient and converges to the global solution more reliably than the state-of-the-art methods.
****
MetaFusion: Infrared and Visible Image Fusion via Meta-Feature Embedding From Object Detection
Wenda Zhao, Shigeng Xie, Fan Zhao, You He, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13955-13965
Fusing infrared and visible images can provide more texture details for subsequent object detection task. Conversely, detection task furnishes object semantic information to improve the infrared and visible image fusion. Thus, a joint fusion and detection learning to use their mutual promotion is attracting more attention. However, the feature gap between these two different-level tasks hinders the progress. Addressing this issue, this paper proposes an infrared and visible image fusion via meta-feature embedding from object detection. The core idea is that meta-feature embedding model is designed to generate object semantic features according to fusion network ability, and thus the semantic features are naturally compatible with fusion features. It is optimized by simulating a meta learning. Moreover, we further implement a mutual promotion learning between fusion and detection tasks to improve their performances. Comprehensive experiments on three public datasets demonstrate the effectiveness of our method. Code and model are available at: https://github.com/wdzhao123/MetaFusion.
****
Spectral Enhanced Rectangle Transformer for Hyperspectral Image Denoising
Miaoyu Li, Ji Liu, Ying Fu, Yulun Zhang, Dejing Dou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5805-5814
Denoising is a crucial step for hyperspectral image (HSI) applications. Though witnessing the great power of deep learning, existing HSI denoising methods suffer from limitations in capturing the non-local self-similarity. Transformers have

 shown potential in capturing long-range dependencies, but few attempts have bee
n made with specifically designed Transformer to model the spatial and spectral
correlation in HSIs. In this paper, we address these issues by proposing a spect
ral enhanced rectangle Transformer, driving it to explore the non-local spatial
similarity and global spectral low-rank property of HSIs. For the former, we exp
loit the rectangle self-attention horizontally and vertically to capture the non
-local similarity in the spatial domain. For the latter, we design a spectral en
hancement module that is capable of extracting global underlying low-rank proper
ty of spatial-spectral cubes to suppress noise, while enabling the interactions
among non-overlapping spatial rectangles. Extensive experiments have been conduc
ted on both synthetic noisy HSIs and real noisy HSIs, showing the effectiveness
of our proposed method in terms of both objective metric and subjective visual q
uality. The code is available at https://github.com/MyuLi/SERT.
****
End-to-End Vectorized HD-Map Construction With Piecewise Bezier Curve
Limeng Qiao, Wenjie Ding, Xi Qiu, Chi Zhang; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13218-13228
Vectorized high-definition map (HD-map) construction, which focuses on the perce
ption of centimeter-level environmental information, has attracted significant r
esearch interest in the autonomous driving community. Most existing approaches f
irst obtain rasterized map with the segmentation-based pipeline and then conduct
 heavy post-processing for downstream-friendly vectorization. In this paper, by
delving into parameterization-based methods, we pioneer a concise and elegant sc
heme that adopts unified piecewise Bezier curve. In order to vectorize changeful
 map elements end-to-end, we elaborate a simple yet effective architecture, name
d Piecewise Bezier HD-map Network (BeMapNet), which is formulated as a direct se
t prediction paradigm and postprocessing-free. Concretely, we first introduce a
novel IPM-PE Align module to inject 3D geometry prior into BEV features through
common position encoding in Transformer. Then a well-designed Piecewise Bezier H
ead is proposed to output the details of each map element, including the coordin
ate of control points and the segment number of curves. In addition, based on th
e progressively restoration of Bezier curve, we also present an efficient Point-
Curve-Region Loss for supervising more robust and precise HD-map modeling. Exten
sive comparisons show that our method is remarkably superior to other existing S
OTAs by 18.0 mAP at least.
****
PointListNet: Deep Learning on 3D Point Lists
Hehe Fan, Linchao Zhu, Yi Yang, Mohan Kankanhalli; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17692-177
01
Deep neural networks on regular 1D lists (e.g., natural languages) and irregular
 3D sets (e.g., point clouds) have made tremendous achievements. The key to natu
ral language processing is to model words and their regular order dependency in
texts. For point cloud understanding, the challenge is to understand the geometr
y via irregular point coordinates, in which point-feeding orders do not matter.
However, there are a few kinds of data that exhibit both regular 1D list and irr
egular 3D set structures, such as proteins and non-coding RNAs. In this paper, w
e refer to them as 3D point lists and propose a Transformer-style PointListNet t
o model them. First, PointListNet employs non-parametric distance-based attentio
n because we find sometimes it is the distance, instead of the feature or type,
that mainly determines how much two points, e.g., amino acids, are correlated in
 the micro world. Second, different from the vanilla Transformer that directly p
erforms a simple linear transformation on inputs to generate values and does not
 explicitly model relative relations, our PointListNet integrates the 1D order a
nd 3D Euclidean displacements into values. We conduct experiments on protein fol
d classification and enzyme reaction classification. Experimental results show t
he effectiveness of the proposed PointListNet.
****
On Data Scaling in Masked Image Modeling
Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Yixuan Wei, Qi Dai, Han Hu; Procee

dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10365-10374

Scaling properties have been one of the central issues in self-supervised pre-training, especially the data scalability, which has successfully motivated the large-scale self-supervised pre-trained language models and endowed them with significant modeling capabilities. However, scaling properties seem to be unintentionally neglected in the recent trending studies on masked image modeling (MIM), and some arguments even suggest that MIM cannot benefit from large-scale data. In this work, we try to break down these preconceptions and systematically study the scaling behaviors of MIM through extensive experiments, with data ranging from 10% of ImageNet-1K to full ImageNet-22K, model parameters ranging from 49-million to one-billion, and training length ranging from 125K to 500K iterations. And our main findings can be summarized in two folds: 1) masked image modeling remains demanding large-scale data in order to scale up computes and model parameters; 2) masked image modeling cannot benefit from more data under a non-overfitting scenario, which diverges from the previous observations in self-supervised pre-trained language models or supervised pre-trained vision models. In addition, we reveal several intriguing properties in MIM, such as high sample efficiency in large MIM models and strong correlation between pre-training validation loss and transfer performance. We hope that our findings could deepen the understanding of masked image modeling and facilitate future developments on large-scale vision models. Code and models will be available at https://github.com/microsoft/SimMIM.

****

Upcycling Models Under Domain and Category Shift

Sanqing Qu, Tianpei Zou, Florian Röhrbein, Cewu Lu, Guang Chen, Dacheng Tao, Changjun Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20019-20028

Deep neural networks (DNNs) often perform poorly in the presence of domain shift and category shift. How to upcycle DNNs and adapt them to the target task remains an important open problem. Unsupervised Domain Adaptation (UDA), especially recently proposed Source-free Domain Adaptation (SFDA), has become a promising technology to address this issue. Nevertheless, most existing SFDA methods require that the source domain and target domain share the same label space, consequently being only applicable to the vanilla closed-set setting. In this paper, we take one step further and explore the Source-free Universal Domain Adaptation (SF-UniDA). The goal is to identify "known" data samples under both domain and category shift, and reject those "unknown" data samples (not present in source classes), with only the knowledge from standard pre-trained source model. To this end, we introduce an innovative global and local clustering learning technique (GLC). Specifically, we design a novel, adaptive one-vs-all global clustering algorithm to achieve the distinction across different target classes and introduce a local k-NN clustering strategy to alleviate negative transfer. We examine the superiority of our GLC on multiple benchmarks with different category shift scenarios, including partial-set, open-set, and open-partial-set DA. More remarkably, in the most challenging open-partial-set DA scenario, GLC outperforms UMAD by 14.8% on the VisDA benchmark.

****

Single Domain Generalization for LiDAR Semantic Segmentation

Hyeonseong Kim, Yoonsu Kang, Changgyoon Oh, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17587-17598

With the success of the 3D deep learning models, various perception technologies for autonomous driving have been developed in the LiDAR domain. While these models perform well in the trained source domain, they struggle in unseen domains with a domain gap. In this paper, we propose a single domain generalization method for LiDAR semantic segmentation (DGLSS) that aims to ensure good performance not only in the source domain but also in the unseen domain by learning only on the source domain. We mainly focus on generalizing from a dense source domain and target the domain shift from different LiDAR sensor configurations and scene di

stributions. To this end, we augment the domain to simulate the unseen domains by randomly subsampling the LiDAR scans. With the augmented domain, we introduce two constraints for generalizable representation learning: sparsity invariant feature consistency (SIFC) and semantic correlation consistency (SCC). The SIFC aligns sparse internal features of the source domain with the augmented domain based on the feature affinity. For SCC, we constrain the correlation between class prototypes to be similar for every LiDAR scan. We also establish a standardized training and evaluation setting for DGLSS. With the proposed evaluation setting, our method showed improved performance in the unseen domains compared to other baselines. Even without access to the target domain, our method performed better than the domain adaptation method. The code is available at https://github.com/gzgzys9887/DGLSS.
****
Balanced Energy Regularization Loss for Out-of-Distribution Detection
Hyunjun Choi, Hawook Jeong, Jin Young Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15691-15700
In the field of out-of-distribution (OOD) detection, a previous method that use auxiliary data as OOD data has shown promising performance. However, the method provides an equal loss to all auxiliary data to differentiate them from inliers. However, based on our observation, in various tasks, there is a general imbalance in the distribution of the auxiliary OOD data across classes. We propose a balanced energy regularization loss that is simple but generally effective for a variety of tasks. Our balanced energy regularization loss utilizes class-wise different prior probabilities for auxiliary data to address the class imbalance in OOD data. The main concept is to regularize auxiliary samples from majority classes, more heavily than those from minority classes. Our approach performs better for OOD detection in semantic segmentation, long-tailed image classification, and image classification than the prior energy regularization loss. Furthermore, our approach achieves state-of-the-art performance in two tasks: OOD detection in semantic segmentation and long-tailed image classification.
****
3D-Aware Face Swapping
Yixuan Li, Chao Ma, Yichao Yan, Wenhan Zhu, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12705-12714
Face swapping is an important research topic in computer vision with wide applications in entertainment and privacy protection. Existing methods directly learn to swap 2D facial images, taking no account of the geometric information of human faces. In the presence of large pose variance between the source and the target faces, there always exist undesirable artifacts on the swapped face. In this paper, we present a novel 3D-aware face swapping method that generates high-fidelity and multi-view-consistent swapped faces from single-view source and target images. To achieve this, we take advantage of the strong geometry and texture prior of 3D human faces, where the 2D faces are projected into the latent space of a 3D generative model. By disentangling the identity and attribute features in the latent space, we succeed in swapping faces in a 3D-aware manner, being robust to pose variations while transferring fine-grained facial details. Extensive experiments demonstrate the superiority of our 3D-aware face swapping framework in terms of visual quality, identity similarity, and multi-view consistency. Code is available at https://lyx0208.github.io/3dSwap.
****
UMat: Uncertainty-Aware Single Image High Resolution Material Capture
Carlos Rodriguez-Pardo, Henar Domínguez-Elvira, David Pascual-Hernández, Elena Garces; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5764-5774
We propose a learning-based method to recover normals, specularity, and roughness from a single diffuse image of a material, using microgeometry appearance as our primary cue. Previous methods that work on single images tend to produce over-smooth outputs with artifacts, operate at limited resolution, or train one model per class with little room for generalization. In contrast, in this work, we p

ropose a novel capture approach that leverages a generative network with attention and a U-Net discriminator, which shows outstanding performance integrating global information at reduced computational complexity. We showcase the performance of our method with a real dataset of digitized textile materials and show that a commodity flatbed scanner can produce the type of diffuse illumination required as input to our method. Additionally, because the problem might be ill-posed --more than a single diffuse image might be needed to disambiguate the specular reflection-- or because the training dataset is not representative enough of the real distribution, we propose a novel framework to quantify the model's confidence about its prediction at test time. Our method is the first one to deal with the problem of modeling uncertainty in material digitization, increasing the trustworthiness of the process and enabling more intelligent strategies for dataset creation, as we demonstrate with an active learning experiment.
****
Similarity Maps for Self-Training Weakly-Supervised Phrase Grounding
Tal Shaharabany, Lior Wolf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6925-6934
A phrase grounding model receives an input image and a text phrase and outputs a suitable localization map. We present an effective way to refine a phrase ground model by considering self-similarity maps extracted from the latent representation of the model's image encoder. Our main insights are that these maps resemble localization maps and that by combining such maps, one can obtain useful pseudo-labels for performing self-training. Our results surpass, by a large margin, the state-of-the-art in weakly supervised phrase grounding. A similar gap in performance is obtained for a recently proposed downstream task called WWbL, in which the input image is given without any text. Our code is available as supplementary.
****
SCOOP: Self-Supervised Correspondence and Optimization-Based Scene Flow
Itai Lang, Dror Aiger, Forrester Cole, Shai Avidan, Michael Rubinstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5281-5290
Scene flow estimation is a long-standing problem in computer vision, where the goal is to find the 3D motion of a scene from its consecutive observations. Recently, there have been efforts to compute the scene flow from 3D point clouds. A common approach is to train a regression model that consumes source and target point clouds and outputs the per-point translation vector. An alternative is to learn point matches between the point clouds concurrently with regressing a refinement of the initial correspondence flow. In both cases, the learning task is very challenging since the flow regression is done in the free 3D space, and a typical solution is to resort to a large annotated synthetic dataset. We introduce SCOOP, a new method for scene flow estimation that can be learned on a small amount of data without employing ground-truth flow supervision. In contrast to previous work, we train a pure correspondence model focused on learning point feature representation and initialize the flow as the difference between a source point and its softly corresponding target point. Then, in the run-time phase, we directly optimize a flow refinement component with a self-supervised objective, which leads to a coherent and accurate flow field between the point clouds. Experiments on widespread datasets demonstrate the performance gains achieved by our method compared to existing leading techniques while using a fraction of the training data. Our code is publicly available.
****
SLACK: Stable Learning of Augmentations With Cold-Start and KL Regularization
Juliette Marrie, Michael Arbel, Diane Larlus, Julien Mairal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24306-24314
Data augmentation is known to improve the generalization capabilities of neural networks, provided that the set of transformations is chosen with care, a selection often performed manually. Automatic data augmentation aims at automating this process. However, most recent approaches still rely on some prior information;

 they start from a small pool of manually-selected default transformations that are either used to pretrain the network or forced to be part of the policy learned by the automatic data augmentation algorithm. In this paper, we propose to directly learn the augmentation policy without leveraging such prior knowledge. The resulting bilevel optimization problem becomes more challenging due to the larger search space and the inherent instability of bilevel optimization algorithms. To mitigate these issues (i) we follow a successive cold-start strategy with a Kullback-Leibler regularization, and (ii) we parameterize magnitudes as continuous distributions. Our approach leads to competitive results on standard benchmarks despite a more challenging setting, and generalizes beyond natural images.
****
Gradient Norm Aware Minimization Seeks First-Order Flatness and Improves Generalization
Xingxuan Zhang, Renzhe Xu, Han Yu, Hao Zou, Peng Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20247-20257
Recently, flat minima are proven to be effective for improving generalization and sharpness-aware minimization (SAM) achieves state-of-the-art performance. Yet the current definition of flatness discussed in SAM and its follow-ups are limited to the zeroth-order flatness (i.e., the worst-case loss within a perturbation radius). We show that the zeroth-order flatness can be insufficient to discriminate minima with low generalization error from those with high generalization error both when there is a single minimum or multiple minima within the given perturbation radius. Thus we present first-order flatness, a stronger measure of flatness focusing on the maximal gradient norm within a perturbation radius which bounds both the maximal eigenvalue of Hessian at local minima and the regularization function of SAM. We also present a novel training procedure named Gradient norm Aware Minimization (GAM) to seek minima with uniformly small curvature across all directions. Experimental results show that GAM improves the generalization of models trained with current optimizers such as SGD and AdamW on various datasets and networks. Furthermore, we show that GAM can help SAM find flatter minima and achieve better generalization.
****
Phone2Proc: Bringing Robust Robots Into Our Chaotic World
Matt Deitke, Rose Hendrix, Ali Farhadi, Kiana Ehsani, Aniruddha Kembhavi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9665-9675
Training embodied agents in simulation has become mainstream for the embodied AI community. However, these agents often struggle when deployed in the physical world due to their inability to generalize to real-world environments. In this paper, we present Phone2Proc, a method that uses a 10-minute phone scan and conditional procedural generation to create a distribution of training scenes that are semantically similar to the target environment. The generated scenes are conditioned on the wall layout and arrangement of large objects from the scan, while also sampling lighting, clutter, surface textures, and instances of smaller objects with randomized placement and materials. Leveraging just a simple RGB camera, training with Phone2Proc shows massive improvements from 34.7% to 70.7% success rate in sim-to-real ObjectNav performance across a test suite of over 200 trials in diverse real-world environments, including homes, offices, and RoboTHOR. Furthermore, Phone2Proc's diverse distribution of generated scenes makes agents remarkably robust to changes in the real world, such as human movement, object rearrangement, lighting changes, or clutter.
****
Latency Matters: Real-Time Action Forecasting Transformer
Harshayu Girase, Nakul Agarwal, Chiho Choi, Karttikeya Mangalam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18759-18769
We present RAFTformer, a real-time action forecasting transformer for latency aware real-world action forecasting applications. RAFTformer is a two-stage fully transformer based architecture which consists of a video transformer backbone th

at operates on high resolution, short range clips and a head transformer encoder that temporally aggregates information from multiple short range clips to span a long-term horizon. Additionally, we propose a self-supervised shuffled causal masking scheme to improve model generalization during training. Finally, we also propose a real-time evaluation setting that directly couples model inference latency to overall forecasting performance and brings forth an hitherto overlooked trade-off between latency and action forecasting performance. Our parsimonious network design facilitates RAFTformer inference latency to be 9x smaller than prior works at the same forecasting accuracy. Owing to its two-staged design, RAFTformer uses 94% less training compute and 90% lesser training parameters to outperform prior state-of-the-art baselines by 4.9 points on EGTEA Gaze+ and by 1.4 points on EPIC-Kitchens-100 dataset, as measured by Top-5 recall (T5R) in the offline setting. In the real-time setting, RAFTformer outperforms prior works by an even greater margin of upto 4.4 T5R points on the EPIC-Kitchens-100 dataset. Project Webpage: https://karttikeya.github.io/publication/RAFTformer/
****
HierVL: Learning Hierarchical Video-Language Embeddings
Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, Kristen Grauman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23066-23078
Video-language embeddings are a promising avenue for injecting semantics into visual representations, but existing methods capture only short-term associations between seconds-long video clips and their accompanying text. We propose HierVL, a novel hierarchical video-language embedding that simultaneously accounts for both long-term and short-term associations. As training data, we take videos accompanied by timestamped text descriptions of human actions, together with a high-level text summary of the activity throughout the long video (as are available in Ego4D). We introduce a hierarchical contrastive training objective that encourages text-visual alignment at both the clip level and video level. While the clip-level constraints use the step-by-step descriptions to capture what is happening in that instant, the video-level constraints use the summary text to capture why it is happening, i.e., the broader context for the activity and the intent of the actor. Our hierarchical scheme yields a clip representation that outperforms its single-level counterpart, as well as a long-term video representation that achieves SotA results on tasks requiring long-term video modeling. HierVL successfully transfers to multiple challenging downstream tasks (in EPIC-KITCHENS-100, Charades-Ego, HowTo100M) in both zero-shot and fine-tuned settings.
****
GraVoS: Voxel Selection for 3D Point-Cloud Detection
Oren Shrout, Yizhak Ben-Shabat, Ayellet Tal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21684-21693
3D object detection within large 3D scenes is challenging not only due to the sparse and irregular 3D point clouds, but also due to both the extreme foreground-background scene imbalance and class imbalance. A common approach is to add ground-truth objects from other scenes. Differently, we propose to modify the scenes by removing elements (voxels), rather than adding ones. Our approach selects the "meaningful" voxels, in a manner that addresses both types of dataset imbalance. The approach is general and can be applied to any voxel-based detector, yet the meaningfulness of a voxel is network-dependent. Our voxel selection is shown to improve the performance of several prominent 3D detection methods.
****
Learning Articulated Shape With Keypoint Pseudo-Labels From Web Images
Anastasis Stathopoulos, Georgios Pavlakos, Ligong Han, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13092-13101
This paper shows that it is possible to learn models for monocular 3D reconstruction of articulated objects (e.g. horses, cows, sheep), using as few as 50-150 images labeled with 2D keypoints. Our proposed approach involves training category-specific keypoint estimators, generating 2D keypoint pseudo-labels on unlabeled web images, and using both the labeled and self-labeled sets to train 3D recon

struction models. It is based on two key insights: (1) 2D keypoint estimation ne
tworks trained on as few as 50-150 images of a given object category generalize
well and generate reliable pseudo-labels; (2) a data selection mechanism can aut
omatically create a "curated" subset of the unlabeled web images that can be use
d for training -- we evaluate four data selection methods. Coupling these two in
sights enables us to train models that effectively utilize web images, resulting
in improved 3D reconstruction performance for several articulated object catego
ries beyond the fully-supervised baseline. Our approach can quickly bootstrap a
model and requires only a few images labeled with 2D keypoints. This requirement
can be easily satisfied for any new object category. To showcase the practicali
ty of our approach for predicting the 3D shape of arbitrary object categories, w
e annotate 2D keypoints on 250 giraffe and bear images from COCO in just 2.5 hou
rs per category.
****
Rethinking Image Super Resolution From Long-Tailed Distribution Learning Perspec
tive
Yuanbiao Gou, Peng Hu, Jiancheng Lv, Hongyuan Zhu, Xi Peng; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
14327-14336
Existing studies have empirically observed that the resolution of the low-freque
ncy region is easier to enhance than that of the high-frequency one. Although pl
entiful works have been devoted to alleviating this problem, little understandin
g is given to explain it. In this paper, we try to give a feasible answer from a
machine learning perspective, i.e., the twin fitting problem caused by the long
-tailed pixel distribution in natural images. With this explanation, we reformul
ate image super resolution (SR) as a long-tailed distribution learning problem a
nd solve it by bridging the gaps of the problem between in low- and high-level v
ision tasks. As a result, we design a long-tailed distribution learning solution
, that rebalances the gradients from the pixels in the low- and high-frequency r
egion, by introducing a static and a learnable structure prior. The learned SR m
odel achieves better balance on the fitting of the low- and high-frequency regio
n so that the overall performance is improved. In the experiments, we evaluate t
he solution on four CNN- and one Transformer-based SR models w.r.t. six datasets
and three tasks, and experimental results demonstrate its superiority.
****
RobustNeRF: Ignoring Distractors With Robust Losses
Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, Andrea
Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 20626-20636
Neural radiance fields (NeRF) excel at synthesizing new views given multi-view,
calibrated images of a static scene. When scenes include distractors, which are
not persistent during image capture (moving objects, lighting variations, shadow
s), artifacts appear as view-dependent effects or 'floaters'. To cope with distr
actors, we advocate a form of robust estimation for NeRF training, modeling dist
ractors in training data as outliers of an optimization problem. Our method succ
essfully removes outliers from a scene and improves upon our baselines, on synth
etic and real-world scenes. Our technique is simple to incorporate in modern NeR
F frameworks, with few hyper-parameters. It does not assume a priori knowledge o
f the types of distractors, and is instead focused on the optimization problem r
ather than pre-processing or modeling transient objects. More results on our pag
e https://robustnerf.github.io/public.
****
Spherical Transformer for LiDAR-Based 3D Recognition
Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, Jiaya Jia; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 175
45-17555
LiDAR-based 3D point cloud recognition has benefited various applications. Witho
ut specially considering the LiDAR point distribution, most current methods suff
er from information disconnection and limited receptive field, especially for th
e sparse distant points. In this work, we study the varying-sparsity distributio

n of LiDAR points and present SphereFormer to directly aggregate information from dense close points to the sparse distant ones. We design radial window self-attention that partitions the space into multiple non-overlapping narrow and long windows. It overcomes the disconnection issue and enlarges the receptive field smoothly and dramatically, which significantly boosts the performance of sparse distant points. Moreover, to fit the narrow and long windows, we propose exponential splitting to yield fine-grained position encoding and dynamic feature selection to increase model representation ability. Notably, our method ranks 1st on both nuScenes and SemanticKITTI semantic segmentation benchmarks with 81.9% and 74.8% mIoU, respectively. Also, we achieve the 3rd place on nuScenes object detection benchmark with 72.8% NDS and 68.5% mAP. Code is available at https://github.com/dvlab-research/SphereFormer.git.
****
Human-Art: A Versatile Human-Centric Dataset Bridging Natural and Artificial Scenes
Xuan Ju, Ailing Zeng, Jianan Wang, Qiang Xu, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 618-629
Humans have long been recorded in a variety of forms since antiquity. For example, sculptures and paintings were the primary media for depicting human beings before the invention of cameras. However, most current human-centric computer vision tasks like human pose estimation and human image generation focus exclusively on natural images in the real world. Artificial humans, such as those in sculptures, paintings, and cartoons, are commonly neglected, making existing models fail in these scenarios. As an abstraction of life, art incorporates humans in both natural and artificial scenes. We take advantage of it and introduce the Human-Art dataset to bridge related tasks in natural and artificial scenarios. Specifically, Human-Art contains 50k high-quality images with over 123k person instances from 5 natural and 15 artificial scenarios, which are annotated with bounding boxes, keypoints, self-contact points, and text information for humans represented in both 2D and 3D. It is, therefore, comprehensive and versatile for various downstream tasks. We also provide a rich set of baseline results and detailed analyses for related tasks, including human detection, 2D and 3D human pose estimation, image generation, and motion transfer. As a challenging dataset, we hope Human-Art can provide insights for relevant research and open up new research questions.
****
Watch or Listen: Robust Audio-Visual Speech Recognition With Visual Corruption Modeling and Reliability Scoring
Joanna Hong, Minsu Kim, Jeongsoo Choi, Yong Man Ro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18783-18794
This paper deals with Audio-Visual Speech Recognition (AVSR) under multimodal input corruption situation where audio inputs and visual inputs are both corrupted, which is not well addressed in previous research directions. Previous studies have focused on how to complement the corrupted audio inputs with the clean visual inputs with the assumption of the availability of clean visual inputs. However, in real life, the clean visual inputs are not always accessible and can even be corrupted by occluded lip region or with noises. Thus, we firstly analyze that the previous AVSR models are not indeed robust to the corruption of multimodal input streams, the audio and the visual inputs, compared to uni-modal models. Then, we design multimodal input corruption modeling to develop robust AVSR models. Lastly, we propose a novel AVSR framework, namely Audio-Visual Reliability Scoring module (AV-RelScore), that is robust to the corrupted multimodal inputs. The AV-RelScore can determine which input modal stream is reliable or not for the prediction and also can exploit the more reliable streams in prediction. The effectiveness of the proposed method is evaluated with comprehensive experiments on popular benchmark databases, LRS2 and LRS3. We also show that the reliability scores obtained by AV-RelScore well reflect the degree of corruption and make the proposed model focus on the reliable multimodal representations.

****
Turning a CLIP Model Into a Scene Text Detector
Wenwen Yu, Yuliang Liu, Wei Hua, Deqiang Jiang, Bo Ren, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6978-6988
The recent large-scale Contrastive Language-Image Pretraining (CLIP) model has shown great potential in various downstream tasks via leveraging the pretrained vision and language knowledge. Scene text, which contains rich textual and visual information, has an inherent connection with a model like CLIP. Recently, pretraining approaches based on vision language models have made effective progresses in the field of text detection. In contrast to these works, this paper proposes a new method, termed TCM, focusing on Turning the CLIP Model directly for text detection without pretraining process. We demonstrate the advantages of the proposed TCM as follows: (1) The underlying principle of our framework can be applied to improve existing scene text detector. (2) It facilitates the few-shot training capability of existing methods, e.g., by using 10% of labeled data, we significantly improve the performance of the baseline method with an average of 22% in terms of the F-measure on 4 benchmarks. (3) By turning the CLIP model into existing scene text detection methods, we further achieve promising domain adaptation ability. The code will be publicly released at https://github.com/wenwenyu/TCM.
****
VisFusion: Visibility-Aware Online 3D Scene Reconstruction From Videos
Huiyu Gao, Wei Mao, Miaomiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17317-17326
We propose VisFusion, a visibility-aware online 3D scene reconstruction approach from posed monocular videos. In particular, we aim to reconstruct the scene from volumetric features. Unlike previous reconstruction methods which aggregate features for each voxel from input views without considering its visibility, we aim to improve the feature fusion by explicitly inferring its visibility from a similarity matrix, computed from its projected features in each image pair. Following previous works, our model is a coarse-to-fine pipeline including a volume sparsification process. Different from their works which sparsify voxels globally with a fixed occupancy threshold, we perform the sparsification on a local feature volume along each visual ray to preserve at least one voxel per ray for more fine details. The sparse local volume is then fused with a global one for online reconstruction. We further propose to predict TSDF in a coarse-to-fine manner by learning its residuals across scales leading to better TSDF predictions. Experimental results on benchmarks show that our method can achieve superior performance with more scene details. Code is available at: https://github.com/huiyu-gao/VisFusion
****
SCOTCH and SODA: A Transformer Video Shadow Detection Framework
Lihao Liu, Jean Prost, Lei Zhu, Nicolas Papadakis, Pietro Liò, Carola-Bibiane Schönlieb, Angelica I. Aviles-Rivero; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10449-10458
Shadows in videos are difficult to detect because of the large shadow deformation between frames. In this work, we argue that accounting for shadow deformation is essential when designing a video shadow detection method. To this end, we introduce the shadow deformation attention trajectory (SODA), a new type of video self-attention module, specially designed to handle the large shadow deformations in videos. Moreover, we present a new shadow contrastive learning mechanism (SCOTCH) which aims at guiding the network to learn a unified shadow representation from massive positive shadow pairs across different videos. We demonstrate empirically the effectiveness of our two contributions in an ablation study. Furthermore, we show that SCOTCH and SODA significantly outperforms existing techniques for video shadow detection. Code is available at the project page: https://lihaoliu-cambridge.github.io/scotch_and_soda/
****
RODIN: A Generative Model for Sculpting 3D Digital Avatars Using Diffusion

Tengfei Wang, Bo Zhang, Ting Zhang, Shuyang Gu, Jianmin Bao, Tadas Baltrusaitis, Jingjing Shen, Dong Chen, Fang Wen, Qifeng Chen, Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4563-4573
This paper presents a 3D diffusion model that automatically generates 3D digital avatars represented as neural radiance fields (NeRFs). A significant challenge for 3D diffusion is that the memory and processing costs are prohibitive for producing high-quality results with rich details. To tackle this problem, we propose the roll-out diffusion network (RODIN), which takes a 3D NeRF model represented as multiple 2D feature maps and rolls out them onto a single 2D feature plane within which we perform 3D-aware diffusion. The RODIN model brings much-needed computational efficiency while preserving the integrity of 3D diffusion by using 3D-aware convolution that attends to projected features in the 2D plane according to their original relationships in 3D. We also use latent conditioning to orchestrate the feature generation with global coherence, leading to high-fidelity avatars and enabling semantic editing based on text prompts. Finally, we use hierarchical synthesis to further enhance details. The 3D avatars generated by our model compare favorably with those produced by existing techniques. We can generate highly detailed avatars with realistic hairstyles and facial hair. We also demonstrate 3D avatar generation from image or text, as well as text-guided editability.
****
On the Pitfall of Mixup for Uncertainty Calibration
Deng-Bao Wang, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, Min-Ling Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7609-7618
By simply taking convex combinations between pairs of samples and their labels, mixup training has been shown to easily improve predictive accuracy. It has been recently found that models trained with mixup also perform well on uncertainty calibration. However, in this study, we found that mixup training usually makes models less calibratable than vanilla empirical risk minimization, which means that it would harm uncertainty estimation when post-hoc calibration is considered. By decomposing the mixup process into data transformation and random perturbation, we suggest that the confidence penalty nature of the data transformation is the reason of calibration degradation. To mitigate this problem, we first investigate the mixup inference strategy and found that despite it improves calibration on mixup, this ensemble-like strategy does not necessarily outperform simple ensemble. Then, we propose a general strategy named mixup inference in training, which adopts a simple decoupling principle for recovering the outputs of raw samples at the end of forward network pass. By embedding the mixup inference, models can be learned from the original one-hot labels and hence avoid the negative impact of confidence penalty. Our experiments show this strategy properly solves mixup's calibration issue without sacrificing the predictive performance, while even improves accuracy than vanilla mixup.
****
Feature Shrinkage Pyramid for Camouflaged Object Detection With Transformers
Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, Huan Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5557-5566
Vision transformers have recently shown strong global context modeling capabilities in camouflaged object detection. However, they suffer from two major limitations: less effective locality modeling and insufficient feature aggregation in decoders, which are not conducive to camouflaged object detection that explores subtle cues from indistinguishable backgrounds. To address these issues, in this paper, we propose a novel transformer-based Feature Shrinkage Pyramid Network (FSPNet), which aims to hierarchically decode locality-enhanced neighboring transformer features through progressive shrinking for camouflaged object detection. Specifically, we propose a non-local token enhancement module (NL-TEM) that employs the non-local mechanism to interact neighboring tokens and explore graph-based high-order relations within tokens to enhance local representations of transfo

rmers. Moreover, we design a feature shrinkage decoder (FSD) with adjacent inter action modules (AIM), which progressively aggregates adjacent transformer featur es through a layer-by-layer shrinkage pyramid to accumulate imperceptible but ef fective cues as much as possible for object information decoding. Extensive quan titative and qualitative experiments demonstrate that the proposed model signifi cantly outperforms the existing 24 competitors on three challenging COD benchmar k datasets under six widely-used evaluation metrics. Our code is publicly availa ble at https://github.com/ZhouHuang23/FSPNet.
****
Matching Is Not Enough: A Two-Stage Framework for Category-Agnostic Pose Estimat ion
Min Shi, Zihao Huang, Xianzheng Ma, Xiaowei Hu, Zhiguo Cao; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7308-7317
Category-agnostic pose estimation (CAPE) aims to predict keypoints for arbitrary categories given support images with keypoint annotations. Existing approaches match the keypoints across the image for localization. However, such a one-stage matching paradigm shows inferior accuracy: the prediction heavily relies on the matching results, which can be noisy due to the open set nature in CAPE. For ex ample, two mirror-symmetric keypoints (e.g., left and right eyes) in the query i mage can both trigger high similarity on certain support keypoints (eyes), which leads to duplicated or opposite predictions. To calibrate the inaccurate matchi ng results, we introduce a two-stage framework, where matched keypoints from the first stage are viewed as similarity-aware position proposals. Then, the model learns to fetch relevant features to correct the initial proposals in the second stage. We instantiate the framework with a transformer model tailored for CAPE. The transformer encoder incorporates specific designs to improve the representa tion and similarity modeling in the first matching stage. In the second stage, s imilarity-aware proposals are packed as queries in the decoder for refinement vi a cross-attention. Our method surpasses the previous best approach by large marg ins on CAPE benchmark MP-100 on both accuracy and efficiency. Code available at https://github.com/flyinglynx/CapeFormer
****
High-Fidelity Guided Image Synthesis With Latent Diffusion Models
Jaskirat Singh, Stephen Gould, Liang Zheng; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5997-6006
Controllable image synthesis with user scribbles has gained huge public interest with the recent advent of text-conditioned latent diffusion models. The user sc ribbles control the color composition while the text prompt provides control ove r the overall image semantics. However, we find that prior works suffer from an intrinsic domain shift problem wherein the generated outputs often lack details and resemble simplistic representations of the target domain. In this paper, we propose a novel guided image synthesis framework, which addresses this problem b y modeling the output image as the solution of a constrained optimization proble m. We show that while computing an exact solution to the optimization is infeasi ble, an approximation of the same can be achieved while just requiring a single pass of the reverse diffusion process. Additionally, we show that by simply defi ning a cross-attention based correspondence between the input text tokens and th e user stroke-painting, the user is also able to control the semantics of differ ent painted regions without requiring any conditional training or finetuning. Hu man user study results show that the proposed approach outperforms the previous state-of-the-art by over 85.32% on the overall user satisfaction scores. Project page for our paper is available at https://1jsingh.github.io/gradop.
****
CodeTalker: Speech-Driven 3D Facial Animation With Discrete Motion Prior
Jinbo Xing, Menghan Xia, Yuechen Zhang, Xiaodong Cun, Jue Wang, Tien-Tsin Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2023, pp. 12780-12790
Speech-driven 3D facial animation has been widely studied, yet there is still a gap to achieving realism and vividness due to the highly ill-posed nature and sc

arcity of audio-visual data. Existing works typically formulate the cross-modal mapping into a regression task, which suffers from the regression-to-mean proble m leading to over-smoothed facial motions. In this paper, we propose to cast spe ech-driven facial animation as a code query task in a finite proxy space of the learned codebook, which effectively promotes the vividness of the generated moti ons by reducing the cross-modal mapping uncertainty. The codebook is learned by self-reconstruction over real facial motions and thus embedded with realistic fa cial motion priors. Over the discrete motion space, a temporal autoregressive mo del is employed to sequentially synthesize facial motions from the input speech signal, which guarantees lip-sync as well as plausible facial expressions. We de monstrate that our approach outperforms current state-of-the-art methods both qu alitatively and quantitatively. Also, a user study further justifies our superio rity in perceptual quality.
****
Towards Transferable Targeted Adversarial Examples
Zhibo Wang, Hongshan Yang, Yunhe Feng, Peng Sun, Hengchang Guo, Zhifei Zhang, Ku i Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2023, pp. 20534-20543
Transferability of adversarial examples is critical for black-box deep learning model attacks. While most existing studies focus on enhancing the transferabilit y of untargeted adversarial attacks, few of them studied how to generate transfe rable targeted adversarial examples that can mislead models into predicting a sp ecific class. Moreover, existing transferable targeted adversarial attacks usual ly fail to sufficiently characterize the target class distribution, thus sufferi ng from limited transferability. In this paper, we propose the Transferable Targ eted Adversarial Attack (TTAA), which can capture the distribution information o f the target class from both label-wise and feature-wise perspectives, to genera te highly transferable targeted adversarial examples. To this end, we design a g enerative adversarial training framework consisting of a generator to produce ta rgeted adversarial examples, and feature-label dual discriminators to distinguis h the generated adversarial examples from the target class images. Specifically, we design the label discriminator to guide the adversarial examples to learn la bel-related distribution information about the target class. Meanwhile, we desig n a feature discriminator, which extracts the feature-wise information with stro ng cross-model consistency, to enable the adversarial examples to learn the tran sferable distribution information. Furthermore, we introduce the random perturba tion dropping to further enhance the transferability by augmenting the diversity of adversarial examples used in the training process. Experiments demonstrate t hat our method achieves excellent performance on the transferability of targeted adversarial examples. The targeted fooling rate reaches 95.13% when transferred from VGG-19 to DenseNet-121, which significantly outperforms the state-of-the-a rt methods.
****
Semi-Supervised Parametric Real-World Image Harmonization
Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, Eli Shechtman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp . 5927-5936
Learning-based image harmonization techniques are usually trained to undo synthe tic global transformations, applied to a masked foreground in a single ground tr uth photo. This simulated data does not model many important appearance mismatch es (illumination, object boundaries, etc.) between foreground and background in real composites, leading to models that do not generalize well and cannot model complex local changes. We propose a new semi-supervised training strategy that a ddresses this problem and lets us learn complex local appearance harmonization f rom unpaired real composites, where foreground and background come from differen t images. Our model is fully parametric. It uses RGB curves to correct the globa l colors and tone and a shading map to model local variations. Our approach outp erforms previous work on established benchmarks and real composites, as shown in a user study, and processes high-resolution images interactively. The code and project page is available at https://kewang0622.github.io/sprih/.

****
C-SFDA: A Curriculum Learning Aided Self-Training Framework for Efficient Source Free Domain Adaptation
Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, Nazanin Rahnavard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24120-24131
Unsupervised domain adaptation (UDA) approaches focus on adapting models trained on a labeled source domain to an unlabeled target domain. In contrast to UDA, source-free domain adaptation (SFDA) is a more practical setup as access to source data is no longer required during adaptation. Recent state-of-the-art (SOTA) methods on SFDA mostly focus on pseudo-label refinement based self-training which generally suffers from two issues: i) inevitable occurrence of noisy pseudo-labels that could lead to early training time memorization, ii) refinement process requires maintaining a memory bank which creates a significant burden in resource constraint scenarios. To address these concerns, we propose C-SFDA, a curriculum learning aided self-training framework for SFDA that adapts efficiently and reliably to changes across domains based on selective pseudo-labeling. Specifically, we employ a curriculum learning scheme to promote learning from a restricted amount of pseudo labels selected based on their reliabilities. This simple yet effective step successfully prevents label noise propagation during different stages of adaptation and eliminates the need for costly memory-bank based label refinement. Our extensive experimental evaluations on both image recognition and semantic segmentation tasks confirm the effectiveness of our method. C-SFDA is also applicable to online test-time domain adaptation and outperforms previous SOTA methods in this task.
****
Learning Visibility Field for Detailed 3D Human Reconstruction and Relighting
Ruichen Zheng, Peng Li, Haoqian Wang, Tao Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 216-226
Detailed 3D reconstruction and photo-realistic relighting of digital humans are essential for various applications. To this end, we propose a novel sparse-view 3d human reconstruction framework that closely incorporates the occupancy field and albedo field with an additional visibility field--it not only resolves occlusion ambiguity in multiview feature aggregation, but can also be used to evaluate light attenuation for self-shadowed relighting. To enhance its training viability and efficiency, we discretize visibility onto a fixed set of sample directions and supply it with coupled geometric 3D depth feature and local 2D image feature. We further propose a novel rendering-inspired loss, namely TransferLoss, to implicitly enforce the alignment between visibility and occupancy field, enabling end-to-end joint training. Results and extensive experiments demonstrate the effectiveness of the proposed method, as it surpasses state-of-the-art in terms of reconstruction accuracy while achieving comparably accurate relighting to ray-traced ground truth.
****
Improving Zero-Shot Generalization and Robustness of Multi-Modal Models
Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, Jiaping Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11093-11101
Multi-modal image-text models such as CLIP and LiT have demonstrated impressive performance on image classification benchmarks and their zero-shot generalization ability is particularly exciting. While the top-5 zero-shot accuracies of these models are very high, the top-1 accuracies are much lower (over 25% gap in some cases). We investigate the reasons for this performance gap and find that many of the failure cases are caused by ambiguity in the text prompts. First, we develop a simple and efficient zero-shot post-hoc method to identify images whose top-1 prediction is likely to be incorrect, by measuring consistency of the predictions w.r.t. multiple prompts and image transformations. We show that our procedure better predicts mistakes, outperforming the popular max logit baseline on selective prediction tasks. Next, we propose a simple and efficient way to improv

e accuracy on such uncertain images by making use of the WordNet hierarchy; specifically we augment the original class by incorporating its parent and children from the semantic label hierarchy, and plug the augmentation into text prompts. We conduct experiments on both CLIP and LiT models with five different ImageNet-based datasets. For CLIP, our method improves the top-1 accuracy by 17.13% on the uncertain subset and 3.6% on the entire ImageNet validation set. We also show that our method improves across ImageNet shifted datasets, four other datasets, and other model architectures such as LiT. Our proposed method is hyperparameter-free, requires no additional model training and can be easily scaled to other large multi-modal architectures. Code is available at https://github.com/gyhandy/Hierarchy-CLIP.
****
Improving Robustness of Vision Transformers by Reducing Sensitivity To Patch Corruptions
Yong Guo, David Stutz, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4108-4118
Despite their success, vision transformers still remain vulnerable to image corruptions, such as noise or blur. Indeed, we find that the vulnerability mainly stems from the unstable self-attention mechanism, which is inherently built upon patch-based inputs and often becomes overly sensitive to the corruptions across patches. For example, when we only occlude a small number of patches with random noise (e.g., 10%), these patch corruptions would lead to severe accuracy drops and greatly distract intermediate attention layers. To address this, we propose a new training method that improves the robustness of transformers from a new perspective -- reducing sensitivity to patch corruptions (RSPC). Specifically, we first identify and occlude/corrupt the most vulnerable patches and then explicitly reduce sensitivity to them by aligning the intermediate features between clean and corrupted examples. We highlight that the construction of patch corruptions is learned adversarially to the following feature alignment process, which is particularly effective and essentially different from existing methods. In experiments, our RSPC greatly improves the stability of attention layers and consistently yields better robustness on various benchmarks, including CIFAR-10/100-C, ImageNet-A, ImageNet-C, and ImageNet-P.
****
VecFontSDF: Learning To Reconstruct and Synthesize High-Quality Vector Fonts via Signed Distance Functions
Zeqing Xia, Bojun Xiong, Zhouhui Lian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1848-1857
Font design is of vital importance in the digital content design and modern printing industry. Developing algorithms capable of automatically synthesizing vector fonts can significantly facilitate the font design process. However, existing methods mainly concentrate on raster image generation, and only a few approaches can directly synthesize vector fonts. This paper proposes an end-to-end trainable method, VecFontSDF, to reconstruct and synthesize high-quality vector fonts using signed distance functions (SDFs). Specifically, based on the proposed SDF-based implicit shape representation, VecFontSDF learns to model each glyph as shape primitives enclosed by several parabolic curves, which can be precisely converted to quadratic Bezier curves that are widely used in vector font products. In this manner, most image generation methods can be easily extended to synthesize vector fonts. Qualitative and quantitative experiments conducted on a publicly-available dataset demonstrate that our method obtains high-quality results on several tasks, including vector font reconstruction, interpolation, and few-shot vector font synthesis, markedly outperforming the state of the art.
****
MSF: Motion-Guided Sequential Fusion for Efficient 3D Object Detection From Point Cloud Sequences
Chenhang He, Ruihuang Li, Yabin Zhang, Shuai Li, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5196-5205
Point cloud sequences are commonly used to accurately detect 3D objects in appli

cations such as autonomous driving. Current top-performing multi-frame detectors mostly follow a Detect-and-Fuse framework, which extracts features from each frame of the sequence and fuses them to detect the objects in the current frame. However, this inevitably leads to redundant computation since adjacent frames are highly correlated. In this paper, we propose an efficient Motion-guided Sequential Fusion (MSF) method, which exploits the continuity of object motion to mine useful sequential contexts for object detection in the current frame. We first generate 3D proposals on the current frame and propagate them to preceding frames based on the estimated velocities. The points-of-interest are then pooled from the sequence and encoded as proposal features. A novel Bidirectional Feature Aggregation (BiFA) module is further proposed to facilitate the interactions of proposal features across frames. Besides, we optimize the point cloud pooling by a voxel-based sampling technique so that millions of points can be processed in several milliseconds. The proposed MSF method achieves not only better efficiency than other multi-frame detectors but also leading accuracy, with 83.12% and 78.30% mAP on the LEVEL1 and LEVEL2 test sets of Waymo Open Dataset, respectively. Codes can be found at https://github.com/skyhehe123/MSF.
****
Modeling the Distributional Uncertainty for Salient Object Detection Models
Xinyu Tian, Jing Zhang, Mochu Xiang, Yuchao Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19660-19670
Most of the existing salient object detection (SOD) models focus on improving the overall model performance, without explicitly explaining the discrepancy between the training and testing distributions. In this paper, we investigate a particular type of epistemic uncertainty, namely distributional uncertainty, for salient object detection. Specifically, for the first time, we explore the existing class-aware distribution gap exploration techniques, i.e. long-tail learning, single-model uncertainty modeling and test-time strategies, and adapt them to model the distributional uncertainty for our class-agnostic task. We define test sample that is dissimilar to the training dataset as being "out-of-distribution" (OOD) samples. Different from the conventional OOD definition, where OOD samples are those not belonging to the closed-world training categories, OOD samples for SOD are those break the basic priors of saliency, i.e. center prior, color contrast prior, compactness prior and etc., indicating OOD as being "continuous" instead of being discrete for our task. We've carried out extensive experimental results to verify effectiveness of existing distribution gap modeling techniques for SOD, and conclude that both train-time single-model uncertainty estimation techniques and weight-regularization solutions that preventing model activation from drifting too much are promising directions for modeling distributional uncertainty for SOD.
****
Kernel Aware Resampler
Michael Bernasconi, Abdelaziz Djelouah, Farnood Salehi, Markus Gross, Christopher Schroers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22347-22355
Deep learning based methods for super-resolution have become state-of-the-art and outperform traditional approaches by a significant margin. From the initial models designed for fixed integer scaling factors (e.g. x2 or x4), efforts were made to explore different directions such as modeling blur kernels or addressing non-integer scaling factors. However, existing works do not provide a sound framework to handle them jointly. In this paper we propose a framework for generic image resampling that not only addresses all the above mentioned issues but extends the sets of possible transforms from upscaling to generic transforms. A key aspect to unlock these capabilities is the faithful modeling of image warping and changes of the sampling rate during the training data preparation. This allows a localized representation of the implicit image degradation that takes into account the reconstruction kernel, the local geometric distortion and the anti-aliasing kernel. Using this spatially variant degradation map as conditioning for our resampling model, we can address with the same model both global transformations, such as upscaling or rotation, and locally varying transformations such lens

distortion or undistortion. Another important contribution is the automatic esti mation of the degradation map in this more complex resampling setting (i.e. blin d image resampling). Finally, we show that state-of-the-art results can be achie ved by predicting kernels to apply on the input image instead of direct color pr ediction. This renders our model applicable for different types of data not seen during the training such as normals.
****
LaserMix for Semi-Supervised LiDAR Semantic Segmentation
Lingdong Kong, Jiawei Ren, Liang Pan, Ziwei Liu; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21705-21715
Densely annotating LiDAR point clouds is costly, which often restrains the scala bility of fully-supervised learning methods. In this work, we study the underexp lored semi-supervised learning (SSL) in LiDAR semantic segmentation. Our core id ea is to leverage the strong spatial cues of LiDAR point clouds to better exploi t unlabeled data. We propose LaserMix to mix laser beams from different LiDAR sc ans and then encourage the model to make consistent and confident predictions be fore and after mixing. Our framework has three appealing properties. 1) Generic: LaserMix is agnostic to LiDAR representations (e.g., range view and voxel), and hence our SSL framework can be universally applied. 2) Statistically grounded: We provide a detailed analysis to theoretically explain the applicability of the proposed framework. 3) Effective: Comprehensive experimental analysis on popula r LiDAR segmentation datasets (nuScenes, SemanticKITTI, and ScribbleKITTI) demon strates our effectiveness and superiority. Notably, we achieve competitive resul ts over fully-supervised counterparts with 2x to 5x fewer labels and improve the supervised-only baseline significantly by relatively 10.8%. We hope this concis e yet high-performing framework could facilitate future research in semi-supervi sed LiDAR segmentation. Code is publicly available.
****
CODA-Prompt: COntinual Decomposed Attention-Based Prompting for Rehearsal-Free C ontinual Learning
James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Don ghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 023, pp. 11909-11919
Computer vision models suffer from a phenomenon known as catastrophic forgetting when learning novel concepts from continuously shifting training data. Typical solutions for this continual learning problem require extensive rehearsal of pre viously seen data, which increases memory costs and may violate data privacy. Re cently, the emergence of large-scale pre-trained vision transformer models has e nabled prompting approaches as an alternative to data-rehearsal. These approache s rely on a key-query mechanism to generate prompts and have been found to be hi ghly resistant to catastrophic forgetting in the well-established rehearsal-free continual learning setting. However, the key mechanism of these methods is not trained end-to-end with the task sequence. Our experiments show that this leads to a reduction in their plasticity, hence sacrificing new task accuracy, and ina bility to benefit from expanded parameter capacity. We instead propose to learn a set of prompt components which are assembled with input-conditioned weights to produce input-conditioned prompts, resulting in a novel attention-based end-to- end key-query scheme. Our experiments show that we outperform the current SOTA m ethod DualPrompt on established benchmarks by as much as 4.5% in average final a ccuracy. We also outperform the state of art by as much as 4.4% accuracy on a co ntinual learning benchmark which contains both class-incremental and domain-incr emental task shifts, corresponding to many practical settings. Our code is avail able at https://github.com/GT-RIPL/CODA-Prompt
****
HypLiLoc: Towards Effective LiDAR Pose Regression With Hyperbolic Fusion
Sijie Wang, Qiyu Kang, Rui She, Wei Wang, Kai Zhao, Yang Song, Wee Peng Tay; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5176-5185
LiDAR relocalization plays a crucial role in many fields, including robotics, au

tonomous driving, and computer vision. LiDAR-based retrieval from a database typically incurs high computation storage costs and can lead to globally inaccurate pose estimations if the database is too sparse. On the other hand, pose regression methods take images or point clouds as inputs and directly regress global poses in an end-to-end manner. They do not perform database matching and are more computationally efficient than retrieval techniques. We propose HypLiLoc, a new model for LiDAR pose regression. We use two branched backbones to extract 3D features and 2D projection features, respectively. We consider multi-modal feature fusion in both Euclidean and hyperbolic spaces to obtain more effective feature representations. Experimental results indicate that HypLiLoc achieves state-of-the-art performance in both outdoor and indoor datasets. We also conduct extensive ablation studies on the framework design, which demonstrate the effectiveness of multi-modal feature extraction and multi-space embedding. Our code is released at: https://github.com/sijieaaa/HypLiLoc
****
Complementary Intrinsics From Neural Radiance Fields and CNNs for Outdoor Scene Relighting
Siqi Yang, Xuanning Cui, Yongjie Zhu, Jiajun Tang, Si Li, Zhaofei Yu, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16600-16609
Relighting an outdoor scene is challenging due to the diverse illuminations and salient cast shadows. Intrinsic image decomposition on outdoor photo collections could partly solve this problem by weakly supervised labels with albedo and normal consistency from multi-view stereo. With neural radiance fields (NeRFs), editing the appearance code could produce more realistic results without explicitly interpreting the outdoor scene image formation. This paper proposes to complement the intrinsic estimation from volume rendering using NeRFs and from inversing the photometric image formation model using convolutional neural networks (CNNs). The former produces richer and more reliable pseudo labels (cast shadows and sky appearances in addition to albedo and normal) for training the latter to predict interpretable and editable lighting parameters via a single-image prediction pipeline. We demonstrate the advantages of our method for both intrinsic image decomposition and relighting for various real outdoor scenes.
****
VideoFusion: Decomposed Diffusion Models for High-Quality Video Generation
Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10209-10218
A diffusion probabilistic model (DPM), which constructs a forward diffusion process by gradually adding noise to data points and learns the reverse denoising process to generate new samples, has been shown to handle complex data distribution. Despite its recent success in image synthesis, applying DPMs to video generation is still challenging due to high-dimensional data spaces. Previous methods usually adopt a standard diffusion process, where frames in the same video clip are destroyed with independent noises, ignoring the content redundancy and temporal correlation. This work presents a decomposed diffusion process via resolving the per-frame noise into a base noise that is shared among all frames and a residual noise that varies along the time axis. The denoising pipeline employs two jointly-learned networks to match the noise decomposition accordingly. Experiments on various datasets confirm that our approach, termed as VideoFusion, surpasses both GAN-based and diffusion-based alternatives in high-quality video generation. We further show that our decomposed formulation can benefit from pre-trained image diffusion models and well-support text-conditioned video creation.
****
Real-Time Multi-Person Eyeblink Detection in the Wild for Untrimmed Video
Wenzheng Zeng, Yang Xiao, Sicheng Wei, Jinfang Gan, Xintao Zhang, Zhiguo Cao, Zhiwen Fang, Joey Tianyi Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13854-13863
Real-time eyeblink detection in the wild can widely serve for fatigue detection, face anti-spoofing, emotion analysis, etc. The existing research efforts genera

lly focus on single-person cases towards trimmed video. However, multi-person sc enario within untrimmed videos is also important for practical applications, whi ch has not been well concerned yet. To address this, we shed light on this resea rch field for the first time with essential contributions on dataset, theory, an d practices. In particular, a large-scale dataset termed MPEblink that involves 686 untrimmed videos with 8748 eyeblink events is proposed under multi-person co nditions. The samples are captured from unconstrained films to reveal "in the wi ld" characteristics. Meanwhile, a real-time multi-person eyeblink detection meth od is also proposed. Being different from the existing counterparts, our proposi tion runs in a one-stage spatio-temporal way with an end-to-end learning capacit y. Specifically, it simultaneously addresses the sub-tasks of face detection, fa ce tracking, and human instance-level eyeblink detection. This paradigm holds 2 main advantages: (1) eyeblink features can be facilitated via the face's global context (e.g., head pose and illumination condition) with joint optimization and interaction, and (2) addressing these sub-tasks in parallel instead of sequenti al manner can save time remarkably to meet the real-time running requirement. Ex periments on MPEblink verify the essential challenges of real-time multi-person eyeblink detection in the wild for untrimmed video. Our method also outperforms existing approaches by large margins and with a high inference speed.
****
Category Query Learning for Human-Object Interaction Classification
Chi Xie, Fangao Zeng, Yue Hu, Shuang Liang, Yichen Wei; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1527 5-15284
Unlike most previous HOI methods that focus on learning better human-object feat ures, we propose a novel and complementary approach called category query learni ng. Such queries are explicitly associated to interaction categories, converted to image specific category representation via a transformer decoder, and learnt via an auxiliary image-level classification task. This idea is motivated by an e arlier multi-label image classification method, but is for the first time applie d for the challenging human-object interaction classification task. Our method i s simple, general and effective. It is validated on three representative HOI bas elines and achieves new state-of-the-art results on two benchmarks.
****
MDQE: Mining Discriminative Query Embeddings To Segment Occluded Instances on Ch allenging Videos
Minghan Li, Shuai Li, Wangmeng Xiang, Lei Zhang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10524-10533
While impressive progress has been achieved, video instance segmentation (VIS) m ethods with per-clip input often fail on challenging videos with occluded object s and crowded scenes. This is mainly because instance queries in these methods c annot encode well the discriminative embeddings of instances, making the query-b ased segmenter difficult to distinguish those 'hard' instances. To address these issues, we propose to mine discriminative query embeddings (MDQE) to segment oc cluded instances on challenging videos. First, we initialize the positional embe ddings and content features of object queries by considering their spatial conte xtual information and the inter-frame object motion. Second, we propose an inter -instance mask repulsion loss to distance each instance from its nearby non-targ et instances. The proposed MDQE is the first VIS method with per-clip input that achieves state-of-the-art results on challenging videos and competitive perform ance on simple videos. In specific, MDQE with ResNet50 achieves 33.0% and 44.5% mask AP on OVIS and YouTube-VIS 2021, respectively. Code of MDQE can be found at https://github.com/MinghanLi/MDQE_CVPR2023.
****
Are We Ready for Vision-Centric Driving Streaming Perception? The ASAP Benchmark
Xiaofeng Wang, Zheng Zhu, Yunpeng Zhang, Guan Huang, Yun Ye, Wenbo Xu, Ziwei Che n, Xingang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2023, pp. 9600-9610
In recent years, vision-centric perception has flourished in various autonomous driving tasks, including 3D detection, semantic map construction, motion forecas

ting, and depth estimation. Nevertheless, the latency of vision-centric approaches is too high for practical deployment (e.g., most camera-based 3D detectors have a runtime greater than 300ms). To bridge the gap between ideal researches and real-world applications, it is necessary to quantify the trade-off between performance and efficiency. Traditionally, autonomous-driving perception benchmarks perform the online evaluation, neglecting the inference time delay. To mitigate the problem, we propose the Autonomous-driving StreAming Perception (ASAP) benchmark, which is the first benchmark to evaluate the online performance of vision-centric perception in autonomous driving. On the basis of the 2Hz annotated nuScenes dataset, we first propose an annotation-extending pipeline to generate high-frame-rate labels for the 12Hz raw images. Referring to the practical deployment, the Streaming Perception Under constRained-computation (SPUR) evaluation protocol is further constructed, where the 12Hz inputs are utilized for streaming evaluation under the constraints of different computational resources. In the ASAP benchmark, comprehensive experiment results reveal that the model rank alters under different constraints, suggesting that the model latency and computation budget should be considered as design choices to optimize the practical deployment. To facilitate further research, we establish baselines for camera-based streaming 3D detection, which consistently enhance the streaming performance across various hardware. The ASAP benchmark will be made publicly available.
****

Robust Model-Based Face Reconstruction Through Weakly-Supervised Outlier Segmentation

Chunlu Li, Andreas Morel-Forster, Thomas Vetter, Bernhard Egger, Adam Kortylewski; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 372-381

In this work, we aim to enhance model-based face reconstruction by avoiding fitting the model to outliers, i.e. regions that cannot be well-expressed by the model such as occluders or make-up. The core challenge for localizing outliers is that they are highly variable and difficult to annotate. To overcome this challenging problem, we introduce a joint Face-autoencoder and outlier segmentation approach (FOCUS).In particular, we exploit the fact that the outliers cannot be fitted well by the face model and hence can be localized well given a high-quality model fitting. The main challenge is that the model fitting and the outlier segmentation are mutually dependent on each other, and need to be inferred jointly. We resolve this chicken-and-egg problem with an EM-type training strategy, where a face autoencoder is trained jointly with an outlier segmentation network. This leads to a synergistic effect, in which the segmentation network prevents the face encoder from fitting to the outliers, enhancing the reconstruction quality. The improved 3D face reconstruction, in turn, enables the segmentation network to better predict the outliers. To resolve the ambiguity between outliers and regions that are difficult to fit, such as eyebrows, we build a statistical prior from synthetic data that measures the systematic bias in model fitting. Experiments on the NoW testset demonstrate that FOCUS achieves SOTA 3D face reconstruction performance among all baselines that are trained without 3D annotation. Moreover, our results on CelebA-HQ and the AR database show that the segmentation network can localize occluders accurately despite being trained without any segmentation annotation.
****

Not All Image Regions Matter: Masked Vector Quantization for Autoregressive Image Generation

Mengqi Huang, Zhendong Mao, Quan Wang, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2002-2011

Existing autoregressive models follow the two-stage generation paradigm that first learns a codebook in the latent space for image reconstruction and then completes the image generation autoregressively based on the learned codebook. However, existing codebook learning simply models all local region information of images without distinguishing their different perceptual importance, which brings redundancy in the learned codebook that not only limits the next stage's autoregre

ssive model's ability to model important structure but also results in high training cost and slow generation speed. In this study, we borrow the idea of importance perception from classical image coding theory and propose a novel two-stage framework, which consists of Masked Quantization VAE (MQ-VAE) and Stackformer, to relieve the model from modeling redundancy. Specifically, MQ-VAE incorporates an adaptive mask module for masking redundant region features before quantization and an adaptive de-mask module for recovering the original grid image feature map to faithfully reconstruct the original images after quantization. Then, Stackformer learns to predict the combination of the next code and its position in the feature map. Comprehensive experiments on various image generation validate our effectiveness and efficiency.
****

Masked Video Distillation: Rethinking Masked Feature Modeling for Self-Supervised Video Representation Learning

Rui Wang, Dongdong Chen, Zuxuan Wu, Yinpeng Chen, Xiyang Dai, Mengchen Liu, Lu Yuan, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6312-6322

Benefiting from masked visual modeling, self-supervised video representation learning has achieved remarkable progress. However, existing methods focus on learning representations from scratch through reconstructing low-level features like raw pixel values. In this paper, we propose masked video distillation (MVD), a simple yet effective two-stage masked feature modeling framework for video representation learning: firstly we pretrain an image (or video) model by recovering low-level features of masked patches, then we use the resulting features as targets for masked feature modeling. For the choice of teacher models, we observe that students taught by video teachers perform better on temporally-heavy video tasks, while image teachers transfer stronger spatial representations for spatially-heavy video tasks. Visualization analysis also indicates different teachers produce different learned patterns for students. To leverage the advantage of different teachers, we design a spatial-temporal co-teaching method for MVD. Specifically, we distill student models from both video teachers and image teachers by masked feature modeling. Extensive experimental results demonstrate that video transformers pretrained with spatial-temporal co-teaching outperform models distilled with a single teacher on a multitude of video datasets. Our MVD with vanilla ViT achieves state-of-the-art performance compared with previous methods on several challenging video downstream tasks. For example, with the ViT-Large model, our MVD achieves 86.4% and 76.7% Top-1 accuracy on Kinetics-400 and Something-Something-v2, outperforming VideoMAE by 1.2% and 2.4% respectively. When a larger ViT-Huge model is adopted, MVD achieves the state-of-the-art performance with 77.3% Top-1 accuracy on Something-Something-v2. Code will be available at https://github.com/ruiwang2021/mvd.
****

Transformer-Based Unified Recognition of Two Hands Manipulating Objects

Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, Seungryul Baek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4769-4778

Understanding the hand-object interactions from an egocentric video has received a great attention recently. So far, most approaches are based on the convolutional neural network (CNN) features combined with the temporal encoding via the long short-term memory (LSTM) or graph convolution network (GCN) to provide the unified understanding of two hands, an object and their interactions. In this paper, we propose the Transformer-based unified framework that provides better understanding of two hands manipulating objects. In our framework, we insert the whole image depicting two hands, an object and their interactions as input and jointly estimate 3 information from each frame: poses of two hands, pose of an object and object types. Afterwards, the action class defined by the hand-object interactions is predicted from the entire video based on the estimated information combined with the contact map that encodes the interaction between two hands and an object. Experiments are conducted on H2O and FPHA benchmark datasets and we demonstrated the superiority of our method achieving the state-of-the-art accuracy

. Ablative studies further demonstrate the effectiveness of each proposed module
.
****
Azimuth Super-Resolution for FMCW Radar in Autonomous Driving
Yu-Jhe Li, Shawn Hunt, Jinhyung Park, Matthew O'Toole, Kris Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17504-17513
We tackle the task of Azimuth (angular dimension) super-resolution for Frequency Modulated Continuous Wave (FMCW) multiple-input multiple-output (MIMO) radar. FMCW MIMO radar is widely used in autonomous driving alongside Lidar and RGB cameras. However, compared to Lidar, MIMO radar is usually of low resolution due to hardware size restrictions. For example, achieving 1-degree azimuth resolution requires at least 100 receivers, but a single MIMO device usually supports at most 12 receivers. Having limitations on the number of receivers is problematic since a high-resolution measurement of azimuth angle is essential for estimating the location and velocity of objects. To improve the azimuth resolution of MIMO radar, we propose a light, yet efficient, Analog-to-Digital super-resolution model (ADC-SR) that predicts or hallucinates additional radar signals using signals from only a few receivers. Compared with the baseline models that are applied to processed radar Range-Azimuth-Doppler (RAD) maps, we show that our ADC-SR method that processes raw ADC signals achieves comparable performance with 98% (50 times) fewer parameters. We also propose a hybrid super-resolution model (Hybrid-SR) combining our ADC-SR with a standard RAD super-resolution model, and show that performance can be improved by a large margin. Experiments on our City-Radar dataset and the RADIal dataset validate the importance of leveraging raw radar ADC signals. To assess the value of our super-resolution model for autonomous driving, we also perform object detection on the results of our super-resolution model and find that our super-resolution model improves detection performance by around 4% in mAP.
****
PDPP:Projected Diffusion for Procedure Planning in Instructional Videos
Hanlin Wang, Yilu Wu, Sheng Guo, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14836-14845
In this paper, we study the problem of procedure planning in instructional videos, which aims to make goal-directed plans given the current visual observations in unstructured real-life videos. Previous works cast this problem as a sequence planning problem and leverage either heavy intermediate visual observations or natural language instructions as supervision, resulting in complex learning schemes and expensive annotation costs. In contrast, we treat this problem as a distribution fitting problem. In this sense, we model the whole intermediate action sequence distribution with a diffusion model (PDPP), and thus transform the planning problem to a sampling process from this distribution. In addition, we remove the expensive intermediate supervision, and simply use task labels from instructional videos as supervision instead. Our model is a U-Net based diffusion model, which directly samples action sequences from the learned distribution with the given start and end observations. Furthermore, we apply an efficient projection method to provide accurate conditional guides for our model during the learning and sampling process. Experiments on three datasets with different scales show that our PDPP model can achieve the state-of-the-art performance on multiple metrics, even without the task supervision. Code and trained models are available at https://github.com/MCG-NJU/PDPP.
****
RangeViT: Towards Vision Transformers for 3D Semantic Segmentation in Autonomous Driving
Angelika Ando, Spyros Gidaris, Andrei Bursuc, Gilles Puy, Alexandre Boulch, Renaud Marlet; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5240-5250
Casting semantic segmentation of outdoor LiDAR point clouds as a 2D problem, e.g., via range projection, is an effective and popular approach. These projection-based methods usually benefit from fast computations and, when combined with tec

hniques which use other point cloud representations, achieve state-of-the-art re
sults. Today, projection-based methods leverage 2D CNNs but recent advances in c
omputer vision show that vision transformers (ViTs) have achieved state-of-the-a
rt results in many image-based benchmarks. In this work, we question if projecti
on-based methods for 3D semantic segmentation can benefit from these latest impr
ovements on ViTs. We answer positively but only after combining them with three
key ingredients: (a) ViTs are notoriously hard to train and require a lot of tra
ining data to learn powerful representations. By preserving the same backbone ar
chitecture as for RGB images, we can exploit the knowledge from long training on
large image collections that are much cheaper to acquire and annotate than poin
t clouds. We reach our best results with pre-trained ViTs on large image dataset
s. (b) We compensate ViTs' lack of inductive bias by substituting a tailored con
volutional stem for the classical linear embedding layer. (c) We refine pixel-wi
se predictions with a convolutional decoder and a skip connection from the convo
lutional stem to combine low-level but fine-grained features of the the convolut
ional stem with the high-level but coarse predictions of the ViT encoder. With t
hese ingredients, we show that our method, called RangeViT, outperforms existing
projection-based methods on nuScenes and SemanticKITTI. The code is available a
t https://github.com/valeoai/rangevit.
****
ProTeGe: Untrimmed Pretraining for Video Temporal Grounding by Video Temporal Gr
ounding
Lan Wang, Gaurav Mittal, Sandra Sajeev, Ye Yu, Matthew Hall, Vishnu Naresh Bodde
ti, Mei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 6575-6585
Video temporal grounding (VTG) is the task of localizing a given natural languag
e text query in an arbitrarily long untrimmed video. While the task involves unt
rimmed videos, all existing VTG methods leverage features from video backbones p
retrained on trimmed videos. This is largely due to the lack of large-scale well
-annotated VTG dataset to perform pretraining. As a result, the pretrained featu
res lack a notion of temporal boundaries leading to the video-text alignment bei
ng less distinguishable between correct and incorrect locations. We present ProT
eGe as the first method to perform VTG-based untrimmed pretraining to bridge the
gap between trimmed pretrained backbones and downstream VTG tasks. ProTeGe reco
nfigures the HowTo100M dataset, with noisily correlated video-text pairs, into a
VTG dataset and introduces a novel Video-Text Similarity-based Grounding Module
and a pretraining objective to make pretraining robust to noise in HowTo100M. E
xtensive experiments on multiple datasets across downstream tasks with all varia
tions of supervision validate that pretrained features from ProTeGe can signific
antly outperform features from trimmed pretrained backbones on VTG.
****
VQACL: A Novel Visual Question Answering Continual Learning Setting
Xi Zhang, Feifei Zhang, Changsheng Xu; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19102-19112
Research on continual learning has recently led to a variety of work in unimodal
community, however little attention has been paid to multimodal tasks like visu
al question answering (VQA). In this paper, we establish a novel VQA Continual L
earning setting named VQACL, which contains two key components: a dual-level tas
k sequence where visual and linguistic data are nested, and a novel composition
testing containing new skill-concept combinations. The former devotes to simulat
ing the ever-changing multimodal datastream in real world and the latter aims at
measuring models' generalizability for cognitive reasoning. Based on our VQACL,
we perform in-depth evaluations of five well-established continual learning met
hods, and observe that they suffer from catastrophic forgetting and have weak ge
neralizability. To address above issues, we propose a novel representation learn
ing method, which leverages a sample-specific and a sample-invariant feature to
learn representations that are both discriminative and generalizable for VQA. Fu
rthermore, by respectively extracting such representation for visual and textual
input, our method can explicitly disentangle the skill and concept. Extensive e
xperimental results illustrate that our method significantly outperforms existin

g models, demonstrating the effectiveness and compositionality of the proposed approach.
****
Efficient Map Sparsification Based on 2D and 3D Discretized Grids
Xiaoyu Zhang, Yun-Hui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12470-12478
Localization in a pre-built map is a basic technique for robot autonomous navigation. Existing mapping and localization methods commonly work well in small-scale environments. As a map grows larger, however, more memory is required and localization becomes inefficient. To solve these problems, map sparsification becomes a practical necessity to acquire a subset of the original map for localization. Previous map sparsification methods add a quadratic term in mixed-integer programming to enforce a uniform distribution of selected landmarks, which requires high memory capacity and heavy computation. In this paper, we formulate map sparsification in an efficient linear form and select uniformly distributed landmarks based on 2D discretized grids. Furthermore, to reduce the influence of different spatial distributions between the mapping and query sequences, which is not considered in previous methods, we also introduce a space constraint term based on 3D discretized grids. The exhaustive experiments in different datasets demonstrate the superiority of the proposed methods in both efficiency and localization performance. The relevant codes will be released at https://github.com/fishmarch/SLAM_Map_Compression.
****
High-Res Facial Appearance Capture From Polarized Smartphone Images
Dejan Azinovi■, Olivier Maury, Christophe Hery, Matthias Nießner, Justus Thies; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16836-16846
We propose a novel method for high-quality facial texture reconstruction from RGB images using a novel capturing routine based on a single smartphone which we equip with an inexpensive polarization foil. Specifically, we turn the flashlight into a polarized light source and add a polarization filter on top of the camera. Leveraging this setup, we capture the face of a subject with cross-polarized and parallel-polarized light. For each subject, we record two short sequences in a dark environment under flash illumination with different light polarization using the modified smartphone. Based on these observations, we reconstruct an explicit surface mesh of the face using structure from motion. We then exploit the camera and light co-location within a differentiable renderer to optimize the facial textures using an analysis-by-synthesis approach. Our method optimizes for high-resolution normal textures, diffuse albedo, and specular albedo using a coarse-to-fine optimization scheme. We show that the optimized textures can be used in a standard rendering pipeline to synthesize high-quality photo-realistic 3D digital humans in novel environments.
****
JAWS: Just a Wild Shot for Cinematic Transfer in Neural Radiance Fields
Xi Wang, Robin Courant, Jinglei Shi, Eric Marchand, Marc Christie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16933-16942
This paper presents JAWS, an optimzation-driven approach that achieves the robust transfer of visual cinematic features from a reference in-the-wild video clip to a newly generated clip. To this end, we rely on an implicit-neural-representation (INR) in a way to compute a clip that shares the same cinematic features as the reference clip. We propose a general formulation of a camera optimization problem in an INR that computes extrinsic and intrinsic camera parameters as well as timing. By leveraging the differentiability of neural representations, we can back-propagate our designed cinematic losses measured on proxy estimators through a NeRF network to the proposed cinematic parameters directly. We also introduce specific enhancements such as guidance maps to improve the overall quality and efficiency. Results display the capacity of our system to replicate well known camera sequences from movies, adapting the framing, camera parameters and timing of the generated video clip to maximize the similarity with the reference cli

p.
****
Class Attention Transfer Based Knowledge Distillation
Ziyao Guo, Haonan Yan, Hui Li, Xiaodong Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11868-11877
Previous knowledge distillation methods have shown their impressive performance on model compression tasks, however, it is hard to explain how the knowledge they transferred helps to improve the performance of the student network. In this work, we focus on proposing a knowledge distillation method that has both high interpretability and competitive performance. We first revisit the structure of mainstream CNN models and reveal that possessing the capacity of identifying class discriminative regions of input is critical for CNN to perform classification. Furthermore, we demonstrate that this capacity can be obtained and enhanced by transferring class activation maps. Based on our findings, we propose class attention transfer based knowledge distillation (CAT-KD). Different from previous KD methods, we explore and present several properties of the knowledge transferred by our method, which not only improve the interpretability of CAT-KD but also contribute to a better understanding of CNN. While having high interpretability, CAT-KD achieves state-of-the-art performance on multiple benchmarks. Code is available at: https://github.com/GzyAftermath/CAT-KD.
****
EfficientSCI: Densely Connected Network With Space-Time Factorization for Large-Scale Video Snapshot Compressive Imaging
Lishun Wang, Miao Cao, Xin Yuan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18477-18486
Video snapshot compressive imaging (SCI) uses a two-dimensional detector to capture consecutive video frames during a single exposure time. Following this, an efficient reconstruction algorithm needs to be designed to reconstruct the desired video frames. Although recent deep learning-based state-of-the-art (SOTA) reconstruction algorithms have achieved good results in most tasks, they still face the following challenges due to excessive model complexity and GPU memory limitations: 1) these models need high computational cost, and 2) they are usually unable to reconstruct large-scale video frames at high compression ratios. To address these issues, we develop an efficient network for video SCI by using dense connections and space-time factorization mechanism within a single residual block, dubbed EfficientSCI. The EfficientSCI network can well establish spatial-temporal correlation by using convolution in the spatial domain and Transformer in the temporal domain, respectively. We are the first time to show that an UHD color video with high compression ratio can be reconstructed from a snapshot 2D measurement using a single end-to-end deep learning model with PSNR above 32 dB. Extensive results on both simulation and real data show that our method significantly outperforms all previous SOTA algorithms with better real-time performance.
****
Exploring Incompatible Knowledge Transfer in Few-Shot Image Generation
Yunqing Zhao, Chao Du, Milad Abdollahzadeh, Tianyu Pang, Min Lin, Shuicheng Yan, Ngai-Man Cheung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7380-7391
Few-shot image generation (FSIG) learns to generate diverse and high-fidelity images from a target domain using a few (e.g., 10) reference samples. Existing FSIG methods select, preserve and transfer prior knowledge from a source generator (pretrained on a related domain) to learn the target generator. In this work, we investigate an underexplored issue in FSIG, dubbed as incompatible knowledge transfer, which would significantly degrade the realisticness of synthetic samples. Empirical observations show that the issue stems from the least significant filters from the source generator. To this end, we propose knowledge truncation to mitigate this issue in FSIG, which is a complementary operation to knowledge preservation and is implemented by a lightweight pruning-based method. Extensive experiments show that knowledge truncation is simple and effective, consistently achieving state-of-the-art performance, including challenging setups where the source and target domains are more distant. Project Page: https://yunqing-me.gith

ub.io/RICK.
****
Temporally Consistent Online Depth Estimation Using Point-Based Fusion
Numair Khan, Eric Penner, Douglas Lanman, Lei Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9119-9129
Depth estimation is an important step in many computer vision problems such as 3D reconstruction, novel view synthesis, and computational photography. Most existing work focuses on depth estimation from single frames. When applied to videos, the result lacks temporal consistency, showing flickering and swimming artifacts. In this paper we aim to estimate temporally consistent depth maps of video streams in an online setting. This is a difficult problem as future frames are not available and the method must choose between enforcing consistency and correcting errors from previous estimations. The presence of dynamic objects further complicates the problem. We propose to address these challenges by using a global point cloud that is dynamically updated each frame, along with a learned fusion approach in image space. Our approach encourages consistency while simultaneously allowing updates to handle errors and dynamic objects. Qualitative and quantitative results show that our method achieves state-of-the-art quality for consistent video depth estimation.
****
Generalizable Implicit Neural Representations via Instance Pattern Composers
Chiheon Kim, Doyup Lee, Saehoon Kim, Minsu Cho, Wook-Shin Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11808-11817
Despite recent advances in implicit neural representations (INRs), it remains challenging for a coordinate-based multi-layer perceptron (MLP) of INRs to learn a common representation across data instances and generalize it for unseen instances. In this work, we introduce a simple yet effective framework for generalizable INRs that enables a coordinate-based MLP to represent complex data instances by modulating only a small set of weights in an early MLP layer as an instance pattern composer; the remaining MLP weights learn pattern composition rules to learn common representations across instances. Our generalizable INR framework is fully compatible with existing meta-learning and hypernetworks in learning to predict the modulated weight for unseen instances. Extensive experiments demonstrate that our method achieves high performance on a wide range of domains such as an audio, image, and 3D object, while the ablation study validates our weight modulation.
****
MotionTrack: Learning Robust Short-Term and Long-Term Motions for Multi-Object Tracking
Zheng Qin, Sanping Zhou, Le Wang, Jinghai Duan, Gang Hua, Wei Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17939-17948
The main challenge of Multi-Object Tracking (MOT) lies in maintaining a continuous trajectory for each target. Existing methods often learn reliable motion patterns to match the same target between adjacent frames and discriminative appearance features to re-identify the lost targets after a long period. However, the reliability of motion prediction and the discriminability of appearances can be easily hurt by dense crowds and extreme occlusions in the tracking process. In this paper, we propose a simple yet effective multi-object tracker, i.e., MotionTrack, which learns robust short-term and long-term motions in a unified framework to associate trajectories from a short to long range. For dense crowds, we design a novel Interaction Module to learn interaction-aware motions from short-term trajectories, which can estimate the complex movement of each target. For extreme occlusions, we build a novel Refind Module to learn reliable long-term motions from the target's history trajectory, which can link the interrupted trajectory with its corresponding detection. Our Interaction Module and Refind Module are embedded in the well-known tracking-by-detection paradigm, which can work in tandem to maintain superior performance. Extensive experimental results on MOT17 a

nd MOT20 datasets demonstrate the superiority of our approach in challenging scenarios, and it achieves state-of-the-art performances at various MOT metrics. We will make the code and trained models publicly available.
****
3D Registration With Maximal Cliques
Xiyu Zhang, Jiaqi Yang, Shikun Zhang, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17745-17754
As a fundamental problem in computer vision, 3D point cloud registration (PCR) aims to seek the optimal pose to align a point cloud pair. In this paper, we present a 3D registration method with maximal cliques (MAC). The key insight is to loosen the previous maximum clique constraint, and to mine more local consensus information in a graph for accurate pose hypotheses generation: 1) A compatibility graph is constructed to render the affinity relationship between initial correspondences. 2) We search for maximal cliques in the graph, each of which represents a consensus set. We perform node-guided clique selection then, where each node corresponds to the maximal clique with the greatest graph weight. 3) Transformation hypotheses are computed for the selected cliques by SVD algorithm and the best hypothesis is used to perform registration. Extensive experiments on U3M, 3DMatch, 3DLoMatch and KITTI demonstrate that MAC effectively increases registration accuracy, outperforms various state-of-the-art methods and boosts the performance of deep-learned methods. MAC combined with deep-learned methods achieves state-of-the-art registration recall of 95.7% / 78.9% on the 3DMatch / 3DLoMatch.
****
What Can Human Sketches Do for Object Detection?
Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15083-15094
Sketches are highly expressive, inherently capturing subjective and fine-grained visual cues. The exploration of such innate properties of human sketches has, however, been limited to that of image retrieval. In this paper, for the first time, we cultivate the expressiveness of sketches but for the fundamental vision task of object detection. The end result is a sketch-enabled object detection framework that detects based on what you sketch -- that "zebra" (e.g., one that is eating the grass) in a herd of zebras (instance-aware detection), and only the part (e.g., "head" of a "zebra") that you desire (part-aware detection). We further dictate that our model works without (i) knowing which category to expect at testing (zero-shot) and (ii) not requiring additional bounding boxes (as per fully supervised) and class labels (as per weakly supervised). Instead of devising a model from the ground up, we show an intuitive synergy between foundation models (e.g., CLIP) and existing sketch models build for sketch-based image retrieval (SBIR), which can already elegantly solve the task -- CLIP to provide model generalisation, and SBIR to bridge the (sketch->photo) gap. In particular, we first perform independent prompting on both sketch and photo branches of an SBIR model to build highly generalisable sketch and photo encoders on the back of the generalisation ability of CLIP. We then devise a training paradigm to adapt the learned encoders for object detection, such that the region embeddings of detected boxes are aligned with the sketch and photo embeddings from SBIR. Evaluating our framework on standard object detection datasets like PASCAL-VOC and MS-COCO outperforms both supervised (SOD) and weakly-supervised object detectors (WSOD) on zero-shot setups. Project Page: https://pinakinathc.github.io/sketch-detect
****
Identity-Preserving Talking Face Generation With Landmark and Appearance Priors
Weizhi Zhong, Chaowei Fang, Yinqi Cai, Pengxu Wei, Gangming Zhao, Liang Lin, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9729-9738
Generating talking face videos from audio attracts lots of research interest. A few person-specific methods can generate vivid videos but require the target speaker's videos for training or fine-tuning. Existing person-generic methods have

difficulty in generating realistic and lip-synced videos while preserving identity information. To tackle this problem, we propose a two-stage framework consisting of audio-to-landmark generation and landmark-to-video rendering procedures. First, we devise a novel Transformer-based landmark generator to infer lip and jaw landmarks from the audio. Prior landmark characteristics of the speaker's face are employed to make the generated landmarks coincide with the facial outline of the speaker. Then, a video rendering model is built to translate the generated landmarks into face images. During this stage, prior appearance information is extracted from the lower-half occluded target face and static reference images, which helps generate realistic and identity-preserving visual content. For effectively exploring the prior information of static reference images, we align static reference images with the target face's pose and expression based on motion fields. Moreover, auditory features are reused to guarantee that the generated face images are well synchronized with the audio. Extensive experiments demonstrate that our method can produce more realistic, lip-synced, and identity-preserving videos than existing person-generic talking face generation methods.
****
All-in-One Image Restoration for Unknown Degradations Using Adaptive Discriminative Filters for Specific Degradations
Dongwon Park, Byung Hyun Lee, Se Young Chun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5815-5824
Image restorations for single degradations have been widely studied, demonstrating excellent performance for each degradation, but can not reflect unpredictable realistic environments with unknown multiple degradations, which may change over time. To mitigate this issue, image restorations for known and unknown multiple degradations have recently been investigated, showing promising results, but require large networks or have sub-optimal architectures for potential interference among different degradations. Here, inspired by the filter attribution integrated gradients (FAIG), we propose an adaptive discriminative filter-based model for specific degradations (ADMS) to restore images with unknown degradations. Our method allows the network to contain degradation-dedicated filters only for about 3% of all network parameters per each degradation and to apply them adaptively via degradation classification (DC) to explicitly disentangle the network for multiple degradations. Our proposed method has demonstrated its effectiveness in comparison studies and achieved state-of-the-art performance in all-in-one image restoration benchmark datasets of both Rain-Noise-Blur and Rain-Snow-Haze.
****
Weakly Supervised Segmentation With Point Annotations for Histopathology Images via Contrast-Based Variational Model
Hongrun Zhang, Liam Burrows, Yanda Meng, Declan Sculthorpe, Abhik Mukherjee, Sarah E. Coupland, Ke Chen, Yalin Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15630-15640
Image segmentation is a fundamental task in the field of imaging and vision. Supervised deep learning for segmentation has achieved unparalleled success when sufficient training data with annotated labels are available. However, annotation is known to be expensive to obtain, especially for histopathology images where the target regions are usually with high morphology variations and irregular shapes. Thus, weakly supervised learning with sparse annotations of points is promising to reduce the annotation workload. In this work, we propose a contrast-based variational model to generate segmentation results, which serve as reliable complementary supervision to train a deep segmentation model for histopathology images. The proposed method considers the common characteristics of target regions in histopathology images and can be trained in an end-to-end manner. It can generate more regionally consistent and smoother boundary segmentation, and is more robust to unlabeled 'novel' regions. Experiments on two different histology datasets demonstrate its effectiveness and efficiency in comparison to previous models. Code is available at: https://github.com/hrzhang1123/CVM_WS_Segmentation.
****
Efficient RGB-T Tracking via Cross-Modality Distillation
Tianlu Zhang, Hongyuan Guo, Qiang Jiao, Qiang Zhang, Jungong Han; Proceedings of

 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 5404-5413
Most current RGB-T trackers adopt a two-stream structure to extract unimodal RGB
 and thermal features and complex fusion strategies to achieve multi-modal featu
re fusion, which require a huge number of parameters, thus hindering their real-
life applications. On the other hand, a compact RGB-T tracker may be computation
ally efficient but encounter non-negligible performance degradation, due to the
weakening of feature representation ability. To remedy this situation, a cross-m
odality distillation framework is presented to bridge the performance gap betwee
n a compact tracker and a powerful tracker. Specifically, a specific-common feat
ure distillation module is proposed to transform the modality-common information
 as well as the modality-specific information from a deeper two-stream network t
o a shallower single-stream network. In addition, a multi-path selection distill
ation module is proposed to instruct a simple fusion module to learn more accura
te multi-modal information from a well-designed fusion mechanism by using multip
le paths. We validate the effectiveness of our method with extensive experiments
 on three RGB-T benchmarks, which achieves state-of-the-art performance but cons
umes much less computational resources.
****
MetaPortrait: Identity-Preserving Talking Head Generation With Fast Personalized
 Adaptation
Bowen Zhang, Chenyang Qi, Pan Zhang, Bo Zhang, HsiangTao Wu, Dong Chen, Qifeng C
hen, Yong Wang, Fang Wen; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2023, pp. 22096-22105
In this work, we propose an ID-preserving talking head generation framework, whi
ch advances previous methods in two aspects. First, as opposed to interpolating
from sparse flow, we claim that dense landmarks are crucial to achieving accurat
e geometry-aware flow fields. Second, inspired by face-swapping methods, we adap
tively fuse the source identity during synthesis, so that the network better pre
serves the key characteristics of the image portrait. Although the proposed mode
l surpasses prior generation fidelity on established benchmarks, personalized fi
ne-tuning is still needed to further make the talking head generation qualified
for real usage. However, this process is rather computationally demanding that i
s unaffordable to standard users. To alleviate this, we propose a fast adaptatio
n model using a meta-learning approach. The learned model can be adapted to a hi
gh-quality personalized model as fast as 30 seconds. Last but not least, a spati
al-temporal enhancement module is proposed to improve the fine details while ens
uring temporal coherency. Extensive experiments prove the significant superiorit
y of our approach over the state of the arts in both one-shot and personalized s
ettings.
****
UniHCP: A Unified Model for Human-Centric Perceptions
Yuanzheng Ci, Yizhou Wang, Meilin Chen, Shixiang Tang, Lei Bai, Feng Zhu, Rui Zh
ao, Fengwei Yu, Donglian Qi, Wanli Ouyang; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17840-17852
Human-centric perceptions (e.g., pose estimation, human parsing, pedestrian dete
ction, person re-identification, etc.) play a key role in industrial application
s of visual models. While specific human-centric tasks have their own relevant s
emantic aspect to focus on, they also share the same underlying semantic structu
re of the human body. However, few works have attempted to exploit such homogene
ity and design a general-propose model for human-centric tasks. In this work, we
 revisit a broad range of human-centric tasks and unify them in a minimalist man
ner. We propose UniHCP, a Unified Model for Human-Centric Perceptions, which uni
fies a wide range of human-centric tasks in a simplified end-to-end manner with
the plain vision transformer architecture. With large-scale joint training on 33
 humancentric datasets, UniHCP can outperform strong baselines on several in-dom
ain and downstream tasks by direct evaluation. When adapted to a specific task,
UniHCP achieves new SOTAs on a wide range of human-centric tasks, e.g., 69.8 mIo
U on CIHP for human parsing, 86.18 mA on PA-100K for attribute prediction, 90.3
mAP on Market1501 for ReID, and 85.8 JI on CrowdHuman for pedestrian detection,

performing better than specialized models tailored for each task. The code and p
retrained model are available at https://github.com/OpenGVLab/UniHCP.
****
Passive Micron-Scale Time-of-Flight With Sunlight Interferometry
Alankar Kotwal, Anat Levin, Ioannis Gkioulekas; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4139-4149
We introduce an interferometric technique for passive time-of-flight imaging and
 depth sensing at micrometer axial resolutions. Our technique uses a full-field
Michelson interferometer, modified to use sunlight as the only light source. The
 large spectral bandwidth of sunlight makes it possible to acquire micrometer-re
solution time-resolved scene responses, through a simple axial scanning operatio
n. Additionally, the angular bandwidth of sunlight makes it possible to capture
time-of-flight measurements insensitive to indirect illumination effects, such a
s interreflections and subsurface scattering. We build an experimental prototype
 that we operate outdoors, under direct sunlight, and in adverse environment con
ditions such as machine vibrations and vehicle traffic. We use this prototype to
 demonstrate, for the first time, passive imaging capabilities such as micromete
r-scale depth sensing robust to indirect illumination, direct-only imaging, and
imaging through diffusers.
****
VoxelNeXt: Fully Sparse VoxelNet for 3D Object Detection and Tracking
Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, Jiaya Jia; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
 pp. 21674-21683
3D object detectors usually rely on hand-crafted proxies, e.g., anchors or cente
rs, and translate well-studied 2D frameworks to 3D. Thus, sparse voxel features
need to be densified and processed by dense prediction heads, which inevitably c
osts extra computation. In this paper, we instead propose VoxelNext for fully sp
arse 3D object detection. Our core insight is to predict objects directly based
on sparse voxel features, without relying on hand-crafted proxies. Our strong sp
arse convolutional network VoxelNeXt detects and tracks 3D objects through voxel
 features entirely. It is an elegant and efficient framework, with no need for s
parse-to-dense conversion or NMS post-processing. Our method achieves a better s
peed-accuracy trade-off than other mainframe detectors on the nuScenes dataset.
For the first time, we show that a fully sparse voxel-based representation works
 decently for LIDAR 3D object detection and tracking. Extensive experiments on n
uScenes, Waymo, and Argoverse2 benchmarks validate the effectiveness of our appr
oach. Without bells and whistles, our model outperforms all existing LIDAR metho
ds on the nuScenes tracking test benchmark.
****
Behavioral Analysis of Vision-and-Language Navigation Agents
Zijiao Yang, Arjun Majumdar, Stefan Lee; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2574-2582
To be successful, Vision-and-Language Navigation (VLN) agents must be able to gr
ound instructions to actions based on their surroundings. In this work, we devel
op a methodology to study agent behavior on a skill-specific basis -- examining
how well existing agents ground instructions about stopping, turning, and moving
 towards specified objects or rooms. Our approach is based on generating skill-s
pecific interventions and measuring changes in agent predictions. We present a d
etailed case study analyzing the behavior of a recent agent and then compare mul
tiple agents in terms of skill-specific competency scores. This analysis suggest
s that biases from training have lasting effects on agent behavior and that exis
ting models are able to ground simple referring expressions. Our comparisons bet
ween models show that skill-specific scores correlate with improvements in overa
ll VLN task performance.
****
Zero-Shot Generative Model Adaptation via Image-Specific Prompt Learning
Jiayi Guo, Chaofei Wang, You Wu, Eric Zhang, Kai Wang, Xingqian Xu, Shiji Song,
Humphrey Shi, Gao Huang; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2023, pp. 11494-11503

Recently, CLIP-guided image synthesis has shown appealing performance on adaptin
g a pre-trained source-domain generator to an unseen target domain. It does not
require any target-domain samples but only the textual domain labels. The traini
ng is highly efficient, e.g., a few minutes. However, existing methods still hav
e some limitations in the quality of generated images and may suffer from the mo
de collapse issue. A key reason is that a fixed adaptation direction is applied
for all cross-domain image pairs, which leads to identical supervision signals.
To address this issue, we propose an Image-specific Prompt Learning (IPL) method
, which learns specific prompt vectors for each source-domain image. This produc
es a more precise adaptation direction for every cross-domain image pair, endowi
ng the target-domain generator with greatly enhanced flexibility. Qualitative an
d quantitative evaluations on various domains demonstrate that IPL effectively i
mproves the quality and diversity of synthesized images and alleviates the mode
collapse. Moreover, IPL is independent of the structure of the generative model,
such as generative adversarial networks or diffusion models. Code is available
at https://github.com/Picsart-AI-Research/IPL-Zero-Shot-Generative-Model-Adaptat
ion.
****
CelebV-Text: A Large-Scale Facial Text-Video Dataset
Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, Wayne Wu; Proce
edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV
PR), 2023, pp. 14805-14814
Text-driven generation models are flourishing in video generation and editing. H
owever, face-centric text-to-video generation remains a challenge due to the lac
k of a suitable dataset containing high-quality videos and highly relevant texts
. This paper presents CelebV-Text, a large-scale, diverse, and high-quality data
set of facial text-video pairs, to facilitate research on facial text-to-video g
eneration tasks. CelebV-Text comprises 70,000 in-the-wild face video clips with
diverse visual content, each paired with 20 texts generated using the proposed s
emi-automatic text generation strategy. The provided texts are of high quality,
describing both static and dynamic attributes precisely. The superiority of Cele
bV-Text over other datasets is demonstrated via comprehensive statistical analys
is of the videos, texts, and text-video relevance. The effectiveness and potenti
al of CelebV-Text are further shown through extensive self-evaluation. A benchma
rk is constructed with representative methods to standardize the evaluation of t
he facial text-to-video generation task. All data and models are publicly availa
ble.
****
Bias in Pruned Vision Models: In-Depth Analysis and Countermeasures
Eugenia Iofinova, Alexandra Peste, Dan Alistarh; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24364-24373
Pruning - that is, setting a significant subset of the parameters of a neural ne
twork to zero - is one of the most popular methods of model compression. Yet, se
veral recent works have raised the issue that pruning may induce or exacerbate b
ias in the output of the compressed model. Despite existing evidence for this ph
enomenon, the relationship between neural network pruning and induced bias is no
t well-understood. In this work, we systematically investigate and characterize
this phenomenon in Convolutional Neural Networks for computer vision. First, we
show that it is in fact possible to obtain highly-sparse models, e.g. with less
than 10% remaining weights, which do not decrease in accuracy nor substantially
increase in bias when compared to dense models. At the same time, we also find t
hat, at higher sparsities, pruned models exhibit higher uncertainty in their out
puts, as well as increased correlations, which we directly link to increased bia
s. We propose easy-to-use criteria which, based only on the uncompressed model,
establish whether bias will increase with pruning, and identify the samples most
susceptible to biased predictions post-compression.
****
AttentionShift: Iteratively Estimated Part-Based Attention Map for Pointly Super
vised Instance Segmentation
Mingxiang Liao, Zonghao Guo, Yuze Wang, Peng Yuan, Bailan Feng, Fang Wan; Procee

dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19519-19528
Pointly supervised instance segmentation (PSIS) learns to segment objects using a single point within the object extent as supervision. Challenged by the non-negligible semantic variance between object parts, however, the single supervision point causes semantic bias and false segmentation. In this study, we propose an AttentionShift method, to solve the semantic bias issue by iteratively decomposing the instance attention map to parts and estimating fine-grained semantics of each part. AttentionShift consists of two modules plugged on the vision transformer backbone: (i) token querying for pointly supervised attention map generation, and (ii) key-point shift, which re-estimates part-based attention maps by key-point filtering in the feature space. These two steps are iteratively performed so that the part-based attention maps are optimized spatially as well as in the feature space to cover full object extent. Experiments on PASCAL VOC and MS COCO 2017 datasets show that AttentionShift respectively improves the state-of-the-art of by 7.7% and 4.8% under mAP@0.5, setting a solid PSIS baseline using vision transformer. Code is enclosed in the supplementary material.
****
Unsupervised Volumetric Animation
Aliaksandr Siarohin, Willi Menapace, Ivan Skorokhodov, Kyle Olszewski, Jian Ren, Hsin-Ying Lee, Menglei Chai, Sergey Tulyakov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4658-4669
We propose a novel approach for unsupervised 3D animation of non-rigid deformable objects. Our method learns the 3D structure and dynamics of objects solely from single-view RGB videos, and can decompose them into semantically meaningful parts that can be tracked and animated. Using a 3D autodecoder framework, paired with a keypoint estimator via a differentiable PnP algorithm, our model learns the underlying object geometry and parts decomposition in an entirely unsupervised manner. This allows it to perform 3D segmentation, 3D keypoint estimation, novel view synthesis, and animation. We primarily evaluate the framework on two video datasets: VoxCeleb 256^2 and TEDXPeople 256^2. In addition, on the Cats 256^2 dataset, we show that it learns compelling 3D geometry even from raw image data. Finally, we show that our model can obtain animatable 3D objects from a singe or a few images.
****
Hard Patches Mining for Masked Image Modeling
Haochen Wang, Kaiyou Song, Junsong Fan, Yuxi Wang, Jin Xie, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10375-10385
Masked image modeling (MIM) has attracted much research attention due to its promising potential for learning scalable visual representations. In typical approaches, models usually focus on predicting specific contents of masked patches, and their performances are highly related to pre-defined mask strategies. Intuitively, this procedure can be considered as training a student (the model) on solving given problems (predict masked patches). However, we argue that the model should not only focus on solving given problems, but also stand in the shoes of a teacher to produce a more challenging problem by itself. To this end, we propose Hard Patches Mining (HPM), a brand-new framework for MIM pre-training. We observe that the reconstruction loss can naturally be the metric of the difficulty of the pre-training task. Therefore, we introduce an auxiliary loss predictor, predicting patch-wise losses first and deciding where to mask next. It adopts a relative relationship learning strategy to prevent overfitting to exact reconstruction loss values. Experiments under various settings demonstrate the effectiveness of HPM in constructing masked images. Furthermore, we empirically find that solely introducing the loss prediction objective leads to powerful representations, verifying the efficacy of the ability to be aware of where is hard to reconstruct.
****
PlaneDepth: Self-Supervised Depth Estimation via Orthogonal Planes
Ruoyu Wang, Zehao Yu, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Co

mputer Vision and Pattern Recognition (CVPR), 2023, pp. 21425-21434
Multiple near frontal-parallel planes based depth representation demonstrated im pressive results in self-supervised monocular depth estimation (MDE). Whereas, s uch a representation would cause the discontinuity of the ground as it is perpen dicular to the frontal-parallel planes, which is detrimental to the identificati on of drivable space in autonomous driving. In this paper, we propose the PlaneD epth, a novel orthogonal planes based presentation, including vertical planes an d ground planes. PlaneDepth estimates the depth distribution using a Laplacian M ixture Model based on orthogonal planes for an input image. These planes are use d to synthesize a reference view to provide the self-supervision signal. Further , we find that the widely used resizing and cropping data augmentation breaks th e orthogonality assumptions, leading to inferior plane predictions. We address t his problem by explicitly constructing the resizing cropping transformation to r ectify the predefined planes and predicted camera pose. Moreover, we propose an augmented self-distillation loss supervised with a bilateral occlusion mask to b oost the robustness of orthogonal planes representation for occlusions. Thanks t o our orthogonal planes representation, we can extract the ground plane in an un supervised manner, which is important for autonomous driving. Extensive experime nts on the KITTI dataset demonstrate the effectiveness and efficiency of our met hod. The code is available at https://github.com/svip-lab/PlaneDepth.
****
Diffusion-SDF: Text-To-Shape via Voxelized Diffusion
Muheng Li, Yueqi Duan, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12642-12651
With the rising industrial attention to 3D virtual modeling technology, generati ng novel 3D content based on specified conditions (e.g. text) has become a hot i ssue. In this paper, we propose a new generative 3D modeling framework called Di ffusion-SDF for the challenging task of text-to-shape synthesis. Previous approa ches lack flexibility in both 3D data representation and shape generation, there by failing to generate highly diversified 3D shapes conforming to the given text descriptions. To address this, we propose a SDF autoencoder together with the V oxelized Diffusion model to learn and generate representations for voxelized sig ned distance fields (SDFs) of 3D shapes. Specifically, we design a novel UinU-Ne t architecture that implants a local-focused inner network inside the standard U -Net architecture, which enables better reconstruction of patch-independent SDF representations. We extend our approach to further text-to-shape tasks including text-conditioned shape completion and manipulation. Experimental results show t hat Diffusion-SDF generates both higher quality and more diversified 3D shapes t hat conform well to given text descriptions when compared to previous approaches . Code is available at: https://github.com/ttlmh/Diffusion-SDF.
****
Compositor: Bottom-Up Clustering and Compositing for Robust Part and Object Segm entation
Ju He, Jieneng Chen, Ming-Xian Lin, Qihang Yu, Alan L. Yuille; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 11259-11268
In this work, we present a robust approach for joint part and object segmentatio n. Specifically, we reformulate object and part segmentation as an optimization problem and build a hierarchical feature representation including pixel, part, a nd object-level embeddings to solve it in a bottom-up clustering manner. Pixels are grouped into several clusters where the part-level embeddings serve as clust er centers. Afterwards, object masks are obtained by compositing the part propos als. This bottom-up interaction is shown to be effective in integrating informat ion from lower semantic levels to higher semantic levels. Based on that, our nov el approach Compositor produces part and object segmentation masks simultaneousl y while improving the mask quality. Compositor achieves state-of-the-art perform ance on PartImageNet and Pascal-Part by outperforming previous methods by around 0.9% and 1.3% on PartImageNet, 0.4% and 1.7% on Pascal-Part in terms of part an d object mIoU and demonstrates better robustness against occlusion by around 4.4 % and 7.1% on part and object respectively.

****
Semantic-Conditional Diffusion Networks for Image Captioning
Jianjie Luo, Yehao Li, Yingwei Pan, Ting Yao, Jianlin Feng, Hongyang Chao, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23359-23368
Recent advances on text-to-image generation have witnessed the rise of diffusion models which act as powerful generative models. Nevertheless, it is not trivial to exploit such latent variable models to capture the dependency among discrete words and meanwhile pursue complex visual-language alignment in image captioning. In this paper, we break the deeply rooted conventions in learning Transformer-based encoder-decoder, and propose a new diffusion model based paradigm tailored for image captioning, namely Semantic-Conditional Diffusion Networks (SCD-Net). Technically, for each input image, we first search the semantically relevant sentences via cross-modal retrieval model to convey the comprehensive semantic information. The rich semantics are further regarded as semantic prior to trigger the learning of Diffusion Transformer, which produces the output sentence in a diffusion process. In SCD-Net, multiple Diffusion Transformer structures are stacked to progressively strengthen the output sentence with better visional-language alignment and linguistical coherence in a cascaded manner. Furthermore, to stabilize the diffusion process, a new self-critical sequence training strategy is designed to guide the learning of SCD-Net with the knowledge of a standard autoregressive Transformer model. Extensive experiments on COCO dataset demonstrate the promising potential of using diffusion models in the challenging image captioning task. Source code is available at https://github.com/YehLi/xmodaler/tree/master/configs/image_caption/scdnet.
****
Unite and Conquer: Plug & Play Multi-Modal Synthesis Using Diffusion Models
Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6070-6079
Generating photos satisfying multiple constraints finds broad utility in the content creation industry. A key hurdle to accomplishing this task is the need for paired data consisting of all modalities (i.e., constraints) and their corresponding output. Moreover, existing methods need retraining using paired data across all modalities to introduce a new condition. This paper proposes a solution to this problem based on denoising diffusion probabilistic models (DDPMs). Our motivation for choosing diffusion models over other generative models comes from the flexible internal structure of diffusion models. Since each sampling step in the DDPM follows a Gaussian distribution, we show that there exists a closed-form solution for generating an image given various constraints. Our method can utilize a single diffusion model trained on multiple sub-tasks and improve the combined task through our proposed sampling strategy. We also introduce a novel reliability parameter that allows using different off-the-shelf diffusion models trained across various datasets during sampling time alone to guide it to the desired outcome satisfying multiple constraints. We perform experiments on various standard multimodal tasks to demonstrate the effectiveness of our approach. More details can be found at: https://nithin-gk.github.io/projectpages/Multidiff
****
TranSG: Transformer-Based Skeleton Graph Prototype Contrastive Learning With Structure-Trajectory Prompted Reconstruction for Person Re-Identification
Haocong Rao, Chunyan Miao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22118-22128
Person re-identification (re-ID) via 3D skeleton data is an emerging topic with prominent advantages. Existing methods usually design skeleton descriptors with raw body joints or perform skeleton sequence representation learning. However, they typically cannot concurrently model different body-component relations, and rarely explore useful semantics from fine-grained representations of body joints. In this paper, we propose a generic Transformer-based Skeleton Graph prototype contrastive learning (TranSG) approach with structure-trajectory prompted reconstruction to fully capture skeletal relations and valuable spatial-temporal sema

ntics from skeleton graphs for person re-ID. Specifically, we first devise the Skeleton Graph Transformer (SGT) to simultaneously learn body and motion relations within skeleton graphs, so as to aggregate key correlative node features into graph representations. Then, we propose the Graph Prototype Contrastive learning (GPC) to mine the most typical graph features (graph prototypes) of each identity, and contrast the inherent similarity between graph representations and different prototypes from both skeleton and sequence levels to learn discriminative graph representations. Last, a graph Structure-Trajectory Prompted Reconstruction (STPR) mechanism is proposed to exploit the spatial and temporal contexts of graph nodes to prompt skeleton graph reconstruction, which facilitates capturing more valuable patterns and graph semantics for person re-ID. Empirical evaluations demonstrate that TranSG significantly outperforms existing state-of-the-art methods. We further show its generality under different graph modeling, RGB-estimated skeletons, and unsupervised scenarios.
****
All Are Worth Words: A ViT Backbone for Diffusion Models
Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22669-22679
Vision transformers (ViT) have shown promise in various vision tasks while the U-Net based on a convolutional neural network (CNN) remains dominant in diffusion models. We design a simple and general ViT-based architecture (named U-ViT) for image generation with diffusion models. U-ViT is characterized by treating all inputs including the time, condition and noisy image patches as tokens and employing long skip connections between shallow and deep layers. We evaluate U-ViT in unconditional and class-conditional image generation, as well as text-to-image generation tasks, where U-ViT is comparable if not superior to a CNN-based U-Net of a similar size. In particular, latent diffusion models with U-ViT achieve record-breaking FID scores of 2.29 in class-conditional image generation on ImageNet 256x256, and 5.48 in text-to-image generation on MS-COCO, among methods without accessing large external datasets during the training of generative models. Our results suggest that, for diffusion-based image modeling, the long skip connection is crucial while the down-sampling and up-sampling operators in CNN-based U-Net are not always necessary. We believe that U-ViT can provide insights for future research on backbones in diffusion models and benefit generative modeling on large scale cross-modality datasets.
****
ZBS: Zero-Shot Background Subtraction via Instance-Level Background Modeling and Foreground Selection
Yongqi An, Xu Zhao, Tao Yu, Haiyun Guo, Chaoyang Zhao, Ming Tang, Jinqiao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6355-6364
Background subtraction (BGS) aims to extract all moving objects in the video frames to obtain binary foreground segmentation masks. Deep learning has been widely used in this field. Compared with supervised-based BGS methods, unsupervised methods have better generalization. However, previous unsupervised deep learning BGS algorithms perform poorly in sophisticated scenarios such as shadows or night lights, and they cannot detect objects outside the pre-defined categories. In this work, we propose an unsupervised BGS algorithm based on zero-shot object detection called Zero-shot Background Subtraction ZBS. The proposed method fully utilizes the advantages of zero-shot object detection to build the open-vocabulary instance-level background model. Based on it, the foreground can be effectively extracted by comparing the detection results of new frames with the background model. ZBS performs well for sophisticated scenarios, and it has rich and extensible categories. Furthermore, our method can easily generalize to other tasks, such as abandoned object detection in unseen environments. We experimentally show that ZBS surpasses state-of-the-art unsupervised BGS methods by 4.70% F-Measure on the CDnet 2014 dataset. The code is released at https://github.com/CASIA-IVA-Lab/ZBS.
****

MobileBrick: Building LEGO for 3D Reconstruction on Mobile Devices
Kejie Li, Jia-Wang Bian, Robert Castle, Philip H.S. Torr, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4892-4901
High-quality 3D ground-truth shapes are critical for 3D object reconstruction evaluation. However, it is difficult to create a replica of an object in reality, and even 3D reconstructions generated by 3D scanners have artefacts that cause biases in evaluation. To address this issue, we introduce a novel multi-view RGBD dataset captured using a mobile device, which includes highly precise 3D ground-truth annotations for 153 object models featuring a diverse set of 3D structures. We obtain precise 3D ground-truth shape without relying on high-end 3D scanners by utilising LEGO models with known geometry as the 3D structures for image capture. The distinct data modality offered by high- resolution RGB images and low-resolution depth maps captured on a mobile device, when combined with precise 3D geometry annotations, presents a unique opportunity for future research on high-fidelity 3D reconstruction. Furthermore, we evaluate a range of 3D reconstruction algorithms on the proposed dataset.
****
GKEAL: Gaussian Kernel Embedded Analytic Learning for Few-Shot Class Incremental Task
Huiping Zhuang, Zhenyu Weng, Run He, Zhiping Lin, Ziqian Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7746-7755
Few-shot class incremental learning (FSCIL) aims to address catastrophic forgetting during class incremental learning in a few-shot learning setting. In this paper, we approach the FSCIL by adopting analytic learning, a technique that converts network training into linear problems. This is inspired by the fact that the recursive implementation (batch-by-batch learning) of analytic learning gives identical weights to that produced by training on the entire dataset at once. The recursive implementation and the weight-identical property highly resemble the FSCIL setting (phase-by-phase learning) and its goal of avoiding catastrophic forgetting. By bridging the FSCIL with the analytic learning, we propose a Gaussian kernel embedded analytic learning (GKEAL) for FSCIL. The key components of GKEAL include the kernel analytic module which allows the GKEAL to conduct FSCIL in a recursive manner, and the augmented feature concatenation module that balances the preference between old and new tasks especially effectively under the few-shot setting. Our experiments show that the GKEAL gives state-of-the-art performance on several benchmark datasets.
****
SteerNeRF: Accelerating NeRF Rendering via Smooth Viewpoint Trajectory
Sicheng Li, Hao Li, Yue Wang, Yiyi Liao, Lu Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20701-20711
Neural Radiance Fields (NeRF) have demonstrated superior novel view synthesis performance but are slow at rendering. To speed up the volume rendering process, many acceleration methods have been proposed at the cost of large memory consumption. To push the frontier of the efficiency-memory trade-off, we explore a new perspective to accelerate NeRF rendering, leveraging a key fact that the viewpoint change is usually smooth and continuous in interactive viewpoint control. This allows us to leverage the information of preceding viewpoints to reduce the number of rendered pixels as well as the number of sampled points along the ray of the remaining pixels. In our pipeline, a low-resolution feature map is rendered first by volume rendering, then a lightweight 2D neural renderer is applied to generate the output image at target resolution leveraging the features of preceding and current frames. We show that the proposed method can achieve competitive rendering quality while reducing the rendering time with little memory overhead, enabling 30FPS at 1080P image resolution with a low memory footprint.
****
Active Exploration of Multimodal Complementarity for Few-Shot Action Recognition
Yuyang Wanyan, Xiaoshan Yang, Chaofan Chen, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6

492-6502
Recently, few-shot action recognition receives increasing attention and achieves remarkable progress. However, previous methods mainly rely on limited unimodal data (e.g., RGB frames) while the multimodal information remains relatively underexplored. In this paper, we propose a novel Active Multimodal Few-shot Action Recognition (AMFAR) framework, which can actively find the reliable modality for each sample based on task-dependent context information to improve few-shot reasoning procedure. In meta-training, we design an Active Sample Selection (ASS) module to organize query samples with large differences in the reliability of modalities into different groups based on modality-specific posterior distributions. In addition, we design an Active Mutual Distillation (AMD) module to capture discriminative task-specific knowledge from the reliable modality to improve the representation learning of unreliable modality by bidirectional knowledge distillation. In meta-test, we adopt Adaptive Multimodal Inference (AMI) module to adaptively fuse the modality-specific posterior distributions with a larger weight on the reliable modality. Extensive experimental results on four public benchmarks demonstrate that our model achieves significant improvements over existing unimodal and multimodal methods.
****
Magic3D: High-Resolution Text-to-3D Content Creation
Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, Tsung-Yi Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 300-309
Recently, DreamFusion demonstrated the utility of a pretrained text-to-image diffusion model to optimize Neural Radiance Fields (NeRF), achieving remarkable text-to-3D synthesis results. However, the method has two inherent limitations: 1) optimization of the NeRF representation is extremely slow, 2) NeRF is supervised by images at a low resolution (64x64), thus leading to low-quality 3D models with a long wait time. In this paper, we address these limitations by utilizing a two-stage coarse-to-fine optimization framework. In the first stage, we use a sparse 3D neural representation to accelerate optimization while using a low-resolution diffusion prior. In the second stage, we use a textured mesh model initialized from the coarse neural representation, allowing us to perform optimization with a very efficient differentiable renderer interacting with high-resolution images. Our method, dubbed Magic3D, can create a 3D mesh model in 40 minutes, 2x faster than DreamFusion (reportedly taking 1.5 hours on average), while achieving 8x higher resolution. User studies show 61.7% raters to prefer our approach than DreamFusion. Together with the image-conditioned generation capabilities, we provide users with new ways to control 3D synthesis, opening up new avenues to various creative applications.
****
Boundary-Aware Backward-Compatible Representation via Adversarial Learning in Image Retrieval
Tan Pan, Furong Xu, Xudong Yang, Sifeng He, Chen Jiang, Qingpei Guo, Feng Qian, Xiaobo Zhang, Yuan Cheng, Lei Yang, Wei Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15201-15210
Image retrieval plays an important role in the Internet world. Usually, the core parts of mainstream visual retrieval systems include an online service of the embedding model and a large-scale vector database. For traditional model upgrades, the old model will not be replaced by the new one until the embeddings of all the images in the database are re-computed by the new model, which takes days or weeks for a large amount of data. Recently, backward-compatible training (BCT) enables the new model to be immediately deployed online by making the new embeddings directly comparable to the old ones. For BCT, improving the compatibility of two models with less negative impact on retrieval performance is the key challenge. In this paper, we introduce AdvBCT, an Adversarial Backward-Compatible Training method with an elastic boundary constraint that takes both compatibility and discrimination into consideration. We first employ adversarial learning to minimize the distribution disparity between embeddings of the new model and the ol

d model. Meanwhile, we add an elastic boundary constraint during training to imp
rove compatibility and discrimination efficiently. Extensive experiments on GLDv
2, Revisited Oxford (ROxford), and Revisited Paris (RParis) demonstrate that our
 method outperforms other BCT methods on both compatibility and discrimination.
The implementation of AdvBCT will be publicly available at https://github.com/As
hespt/AdvBCT.
****
Spatial-Frequency Mutual Learning for Face Super-Resolution
Chenyang Wang, Junjun Jiang, Zhiwei Zhong, Xianming Liu; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 223
56-22366
Face super-resolution (FSR) aims to reconstruct high-resolution (HR) face images
 from the low-resolution (LR) ones. With the advent of deep learning, the FSR te
chnique has achieved significant breakthroughs. However, existing FSR methods ei
ther have a fixed receptive field or fail to maintain facial structure, limiting
 the FSR performance. To circumvent this problem, Fourier transform is introduce
d, which can capture global facial structure information and achieve image-size
receptive field. Relying on the Fourier transform, we devise a spatial-frequency
 mutual network (SFMNet) for FSR, which is the first FSR method to explore the c
orrelations between spatial and frequency domains as far as we know. To be speci
fic, our SFMNet is a two-branch network equipped with a spatial branch and a fre
quency branch. Benefiting from the property of Fourier transform, the frequency
branch can achieve image-size receptive field and capture global dependency whil
e the spatial branch can extract local dependency. Considering that these depend
encies are complementary and both favorable for FSR, we further develop a freque
ncy-spatial interaction block (FSIB) which mutually amalgamates the complementar
y spatial and frequency information to enhance the capability of the model. Quan
titative and qualitative experimental results show that the proposed method outp
erforms state-of-the-art FSR methods in recovering face images. The implementati
on and model will be released at https://github.com/wcy-cs/SFMNet.
****
Sketch2Saliency: Learning To Detect Salient Objects From Human Drawings
Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath C
howdhury, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2023, pp. 2733-2743
Human sketch has already proved its worth in various visual understanding tasks
(e.g., retrieval, segmentation, image-captioning, etc). In this paper, we reveal
 a new trait of sketches -- that they are also salient. This is intuitive as ske
tching is a natural attentive process at its core. More specifically, we aim to
study how sketches can be used as a weak label to detect salient objects present
 in an image. To this end, we propose a novel method that emphasises on how "sal
ient object" could be explained by hand-drawn sketches. To accomplish this, we i
ntroduce a photo-to-sketch generation model that aims to generate sequential ske
tch coordinates corresponding to a given visual photo through a 2D attention mec
hanism. Attention maps accumulated across the time steps give rise to salient re
gions in the process. Extensive quantitative and qualitative experiments prove o
ur hypothesis and delineate how our sketch-based saliency detection model gives
a competitive performance compared to the state-of-the-art.
****
Efficient Frequency Domain-Based Transformers for High-Quality Image Deblurring
Lingshun Kong, Jiangxin Dong, Jianjun Ge, Mingqiang Li, Jinshan Pan; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 5886-5895
We present an effective and efficient method that explores the properties of Tra
nsformers in the frequency domain for high-quality image deblurring. Our method
is motivated by the convolution theorem that the correlation or convolution of t
wo signals in the spatial domain is equivalent to an element-wise product of the
m in the frequency domain. This inspires us to develop an efficient frequency do
main-based self-attention solver (FSAS) to estimate the scaled dot-product atten
tion by an element-wise product operation instead of the matrix multiplication i

n the spatial domain. In addition, we note that simply using the naive feed-forward network (FFN) in Transformers does not generate good deblurred results. To overcome this problem, we propose a simple yet effective discriminative frequency domain-based FFN (DFFN), where we introduce a gated mechanism in the FFN based on the Joint Photographic Experts Group (JPEG) compression algorithm to discriminatively determine which low- and high-frequency information of the features should be preserved for latent clear image restoration. We formulate the proposed FSAS and DFFN into an asymmetrical network based on an encoder and decoder architecture, where the FSAS is only used in the decoder module for better image deblurring. Experimental results show that the proposed method performs favorably against the state-of-the-art approaches.
****
Distilling Focal Knowledge From Imperfect Expert for 3D Object Detection
Jia Zeng, Li Chen, Hanming Deng, Lewei Lu, Junchi Yan, Yu Qiao, Hongyang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 992-1001
Multi-camera 3D object detection blossoms in recent years and most of state-of-the-art methods are built up on the bird's-eye-view (BEV) representations. Albeit remarkable performance, these works suffer from low efficiency. Typically, knowledge distillation can be used for model compression. However, due to unclear 3D geometry reasoning, expert features usually contain some noisy and confusing areas. In this work, we investigate on how to distill the knowledge from an imperfect expert. We propose FD3D, a Focal Distiller for 3D object detection. Specifically, a set of queries are leveraged to locate the instance-level areas for masked feature generation, to intensify feature representation ability in these areas. Moreover, these queries search out the representative fine-grained positions for refined distillation. We verify the effectiveness of our method by applying it to two popular detection models, BEVFormer and DETR3D. The results demonstrate that our method achieves improvements of 4.07 and 3.17 points respectively in terms of NDS metric on nuScenes benchmark. Code is hosted at https://github.com/OpenPerceptionX/BEVPerception-Survey-Recipe.
****
ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding
Le Xue, Mingfei Gao, Chen Xing, 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), 2023, pp. 1179-1189
The recognition capabilities of current state-of-the-art 3D models are limited by datasets with a small number of annotated data and a pre-defined set of categories. In its 2D counterpart, recent advances have shown that similar problems can be significantly alleviated by employing knowledge from other modalities, such as language. Inspired by this, leveraging multimodal information for 3D modality could be promising to improve 3D understanding under the restricted data regime, but this line of research is not well studied. Therefore, we introduce ULIP to learn a unified representation of images, language, and 3D point clouds by pre-training with object triplets from the three modalities. To overcome the shortage of training triplets, ULIP leverages a pre-trained vision-language model that has already learned a common visual and textual space by training with massive image-text pairs. Then, ULIP learns a 3D representation space aligned with the common image-text space, using a small number of automatically synthesized triplets. ULIP is agnostic to 3D backbone networks and can easily be integrated into any 3D architecture. Experiments show that ULIP effectively improves the performance of multiple recent 3D backbones by simply pre-training them on ShapeNet55 using our framework, achieving state-of-the-art performance in both standard 3D classification and zero-shot 3D classification on ModelNet40 and ScanObjectNN. ULIP also improves the performance of PointMLP by around 3% in 3D classification on ScanObjectNN, and outperforms PointCLIP by 28.8% on top-1 accuracy for zero-shot 3D classification on ModelNet40. Our code and pre-trained models are released at https://github.com/salesforce/ULIP.
****

Being Comes From Not-Being: Open-Vocabulary Text-to-Motion Generation With Wordless Training
Junfan Lin, Jianlong Chang, Lingbo Liu, Guanbin Li, Liang Lin, Qi Tian, Chang-Wen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23222-23231
Text-to-motion generation is an emerging and challenging problem, which aims to synthesize motion with the same semantics as the input text. However, due to the lack of diverse labeled training data, most approaches either limit to specific types of text annotations or require online optimizations to cater to the texts during inference at the cost of efficiency and stability. In this paper, we investigate offline open-vocabulary text-to-motion generation in a zero-shot learning manner that neither requires paired training data nor extra online optimization to adapt for unseen texts. Inspired by the prompt learning in NLP, we pretrain a motion generator that learns to reconstruct the full motion from the masked motion. During inference, instead of changing the motion generator, our method reformulates the input text into a masked motion as the prompt for the motion generator to "reconstruct" the motion. In constructing the prompt, the unmasked poses of the prompt are synthesized by a text-to-pose generator. To supervise the optimization of the text-to-pose generator, we propose the first text-pose alignment model for measuring the alignment between texts and 3D poses. And to prevent the pose generator from overfitting to limited training texts, we further propose a novel wordless training mechanism that optimizes the text-to-pose generator without any training texts. The comprehensive experimental results show that our method obtains a significant improvement against the baseline methods. The code is available at https://github.com/junfanlin/oohmg.
****
Deep Learning of Partial Graph Matching via Differentiable Top-K
Runzhong Wang, Ziao Guo, Shaofei Jiang, Xiaokang Yang, Junchi Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6272-6281
Graph matching (GM) aims at discovering node matching between graphs, by maximizing the node- and edge-wise affinities between the matched elements. As an NP-hard problem, its challenge is further pronounced in the existence of outlier nodes in both graphs which is ubiquitous in practice, especially for vision problems. However, popular affinity-maximization-based paradigms often lack a principled scheme to suppress the false matching and resort to handcrafted thresholding to dismiss the outliers. This limitation is also inherited by the neural GM solvers though they have shown superior performance in the ideal no-outlier setting. In this paper, we propose to formulate the partial GM problem as the top-k selection task with a given/estimated number of inliers k. Specifically, we devise a differentiable top-k module that enables effective gradient descent over the optimal-transport layer, which can be readily plugged into SOTA deep GM pipelines including the quadratic matching network NGMv2 as well as the linear matching network GCAN. Meanwhile, the attention-fused aggregation layers are developed to estimate k to enable automatic outlier-robust matching in the wild. Last but not least, we remake and release a new benchmark called IMC-PT-SparseGM, originating from the IMC-PT stereo-matching dataset. The new benchmark involves more scale-varying graphs and partial matching instances from the real world. Experiments show that our methods outperform other partial matching schemes on popular benchmarks.
****
Super-CLEVR: A Virtual Benchmark To Diagnose Domain Robustness in Visual Reasoning
Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, Alan L. Yuille; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14963-14973
Visual Question Answering (VQA) models often perform poorly on out-of-distribution data and struggle on domain generalization. Due to the multi-modal nature of this task, multiple factors of variation are intertwined, making generalization difficult to analyze. This motivates us to introduce a virtual benchmark, Super-

CLEVR, where different factors in VQA domain shifts can be isolated in order that their effects can be studied independently. Four factors are considered: visual complexity, question redundancy, concept distribution and concept compositionality. With controllably generated data, Super-CLEVR enables us to test VQA methods in situations where the test data differs from the training data along each of these axes. We study four existing methods, including two neural symbolic methods NSCL and NSVQA, and two non-symbolic methods FiLM and mDETR; and our proposed method, probabilistic NSVQA (P-NSVQA), which extends NSVQA with uncertainty reasoning. P-NSVQA outperforms other methods on three of the four domain shift factors. Our results suggest that disentangling reasoning and perception, combined with probabilistic uncertainty, form a strong VQA model that is more robust to domain shifts. The dataset and code are released at https://github.com/Lizw14/Super-CLEVR.
****
MonoHuman: Animatable Human Neural Field From Monocular Video
Zhengming Yu, Wei Cheng, Xian Liu, Wayne Wu, Kwan-Yee Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16943-16953
Animating virtual avatars with free-view control is crucial for various applications like virtual reality and digital entertainment. Previous studies have attempted to utilize the representation power of the neural radiance field (NeRF) to reconstruct the human body from monocular videos. Recent works propose to graft a deformation network into the NeRF to further model the dynamics of the human neural field for animating vivid human motions. However, such pipelines either rely on pose-dependent representations or fall short of motion coherency due to frame-independent optimization, making it difficult to generalize to unseen pose sequences realistically. In this paper, we propose a novel framework MonoHuman, which robustly renders view-consistent and high-fidelity avatars under arbitrary novel poses. Our key insight is to model the deformation field with bi-directional constraints and explicitly leverage the off-the-peg keyframe information to reason the feature correlations for coherent results. Specifically, we first propose a Shared Bidirectional Deformation module, which creates a pose-independent generalizable deformation field by disentangling backward and forward deformation correspondences into shared skeletal motion weight and separate non-rigid motions. Then, we devise a Forward Correspondence Search module, which queries the correspondence feature of keyframes to guide the rendering network. The rendered results are thus multi-view consistent with high fidelity, even under challenging novel pose settings. Extensive experiments demonstrate the superiority of our proposed MonoHuman over state-of-the-art methods.
****
Sliced Optimal Partial Transport
Yikun Bai, Bernhard Schmitzer, Matthew Thorpe, Soheil Kolouri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13681-13690
Optimal transport (OT) has become exceedingly popular in machine learning, data science, and computer vision. The core assumption in the OT problem is the equal total amount of mass in source and target measures, which limits its application. Optimal Partial Transport (OPT) is a recently proposed solution to this limitation. Similar to the OT problem, the computation of OPT relies on solving a linear programming problem (often in high dimensions), which can become computationally prohibitive. In this paper, we propose an efficient algorithm for calculating the OPT problem between two non-negative measures in one dimension. Next, following the idea of sliced OT distances, we utilize slicing to define the sliced OPT distance. Finally, we demonstrate the computational and accuracy benefits of the sliced OPT-based method in various numerical experiments. In particular, we show an application of our proposed Sliced-OPT in noisy point cloud registration.
****
Siamese DETR
Zeren Chen, Gengshi Huang, Wei Li, Jianing Teng, Kun Wang, Jing Shao, Chen Chang

e Loy, Lu Sheng; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2023, pp. 15722-15731
Recent self-supervised methods are mainly designed for representation learning w
ith the base model, e.g., ResNets or ViTs. They cannot be easily transferred to
DETR, with task-specific Transformer modules. In this work, we present Siamese D
ETR, a Siamese self-supervised pretraining approach for the Transformer architec
ture in DETR. We consider learning view-invariant and detection-oriented represe
ntations simultaneously through two complementary tasks, i.e., localization and
discrimination, in a novel multi-view learning framework. Two self-supervised pr
etext tasks are designed: (i) Multi-View Region Detection aims at learning to lo
calize regions-of-interest between augmented views of the input, and (ii) Multi-
View Semantic Discrimination attempts to improve object-level discrimination for
 each region. The proposed Siamese DETR achieves state-of-the-art transfer perfo
rmance on COCO and PASCAL VOC detection using different DETR variants in all set
ups. Code is available at https://github.com/Zx55/SiameseDETR.
****
SINE: Semantic-Driven Image-Based NeRF Editing With Prior-Guided Editing Field
Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guo
feng Zhang, Zhaopeng Cui; Proceedings of the IEEE/CVF Conference on Computer Vis
ion and Pattern Recognition (CVPR), 2023, pp. 20919-20929
Despite the great success in 2D editing using user-friendly tools, such as Photo
shop, semantic strokes, or even text prompts, similar capabilities in 3D areas a
re still limited, either relying on 3D modeling skills or allowing editing withi
n only a few categories. In this paper, we present a novel semantic-driven NeRF
editing approach, which enables users to edit a neural radiance field with a sin
gle image, and faithfully delivers edited novel views with high fidelity and mul
ti-view consistency. To achieve this goal, we propose a prior-guided editing fie
ld to encode fine-grained geometric and texture editing in 3D space, and develop
 a series of techniques to aid the editing process, including cyclic constraints
 with a proxy mesh to facilitate geometric supervision, a color compositing mech
anism to stabilize semantic-driven texture editing, and a feature-cluster-based
regularization to preserve the irrelevant content unchanged. Extensive experimen
ts and editing examples on both real-world and synthetic data demonstrate that o
ur method achieves photo-realistic 3D editing using only a single edited image,
pushing the bound of semantic-driven editing in 3D real-world scenes.
****
Turning Strengths Into Weaknesses: A Certified Robustness Inspired Attack Framew
ork Against Graph Neural Networks
Binghui Wang, Meng Pang, Yun Dong; Proceedings of the IEEE/CVF Conference on Com
puter Vision and Pattern Recognition (CVPR), 2023, pp. 16394-16403
Graph neural networks (GNNs) have achieved state-of-the-art performance in many
graph-related tasks such as node classification. However, recent studies show th
at GNNs are vulnerable to both test-time and training-time attacks that perturb
the graph structure. While the existing attack methods have shown promising atta
ck performance, we would like to design an attack framework that can significant
ly enhance both the existing evasion and poisoning attacks. In particular, our a
ttack framework is inspired by certified robustness. Certified robustness was or
iginally used by defenders to defend against adversarial attacks. We are the fir
st, from the attacker perspective, to leverage its properties to better attack G
NNs. Specifically, we first leverage and derive nodes' certified perturbation si
zes against evasion and poisoning attacks based on randomized smoothing. A large
r certified perturbation size of a node indicates this node is theoretically mor
e robust to graph perturbations. Such a property motivates us to focus more on n
odes with smaller certified perturbation sizes, as they are easier to be attacke
d after graph perturbations. Accordingly, we design a certified robustness inspi
red attack loss, when incorporated into (any) existing attacks, produces our cer
tified robustness inspired attack framework. We apply our attack framework to th
e existing attacks and results show it can significantly enhance the existing at
tacks' performance.
****

Demystifying Causal Features on Adversarial Examples and Causal Inoculation for Robust Network by Adversarial Instrumental Variable Regression
Junho Kim, Byung-Kwan Lee, Yong Man Ro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12302-12312
The origin of adversarial examples is still inexplicable in research fields, and it arouses arguments from various viewpoints, albeit comprehensive investigations. In this paper, we propose a way of delving into the unexpected vulnerability in adversarially trained networks from a causal perspective, namely adversarial instrumental variable (IV) regression. By deploying it, we estimate the causal relation of adversarial prediction under an unbiased environment dissociated from unknown confounders. Our approach aims to demystify inherent causal features on adversarial examples by leveraging a zero-sum optimization game between a causal feature estimator (i.e., hypothesis model) and worst-case counterfactuals (i.e., test function) disturbing to find causal features. Through extensive analyses, we demonstrate that the estimated causal features are highly related to the correct prediction for adversarial robustness, and the counterfactuals exhibit extreme features significantly deviating from the correct prediction. In addition, we present how to effectively inoculate CAusal FEatures (CAFE) into defense networks for improving adversarial robustness.
****
NVTC: Nonlinear Vector Transform Coding
Runsen Feng, Zongyu Guo, Weiping Li, Zhibo Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6101-6110
In theory, vector quantization (VQ) is always better than scalar quantization (SQ) in terms of rate-distortion (R-D) performance. Recent state-of-the-art methods for neural image compression are mainly based on nonlinear transform coding (NTC) with uniform scalar quantization, overlooking the benefits of VQ due to its exponentially increased complexity. In this paper, we first investigate on some toy sources, demonstrating that even if modern neural networks considerably enhance the compression performance of SQ with nonlinear transform, there is still an insurmountable chasm between SQ and VQ. Therefore, revolving around VQ, we propose a novel framework for neural image compression named Nonlinear Vector Transform Coding (NVTC). NVTC solves the critical complexity issue of VQ through (1) a multi-stage quantization strategy and (2) nonlinear vector transforms. In addition, we apply entropy-constrained VQ in latent space to adaptively determine the quantization boundaries for joint rate-distortion optimization, which improves the performance both theoretically and experimentally. Compared to previous NTC approaches, NVTC demonstrates superior rate-distortion performance, faster decoding speed, and smaller model size. Our code is available at https://github.com/USTC-IMCL/NVTC.
****
B-Spline Texture Coefficients Estimator for Screen Content Image Super-Resolution
Byeonghyun Pak, Jaewon Lee, Kyong Hwan Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10062-10071
Screen content images (SCIs) include many informative components, e.g., texts and graphics. Such content creates sharp edges or homogeneous areas, making a pixel distribution of SCI different from the natural image. Therefore, we need to properly handle the edges and textures to minimize information distortion of the contents when a display device's resolution differs from SCIs. To achieve this goal, we propose an implicit neural representation using B-splines for screen content image super-resolution (SCI SR) with arbitrary scales. Our method extracts scaling, translating, and smoothing parameters of B-splines. The followed multi-layer perceptron (MLP) uses the estimated B-splines to recover high-resolution SCI. Our network outperforms both a transformer-based reconstruction and an implicit Fourier representation method in almost upscaling factor, thanks to the positive constraint and compact support of the B-spline basis. Moreover, our SR results are recognized as correct text letters with the highest confidence by a pre-trained scene text recognition network. Source code is available at https://github.com/ByeongHyunPak/btc.

****
MetaCLUE: Towards Comprehensive Visual Metaphors Research

Arjun R. Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T. Freeman, Yuanzhen Li, Varun Jampani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23201-23211

Creativity is an indispensable part of human cognition and also an inherent part of how we make sense of the world. Metaphorical abstraction is fundamental in communicating creative ideas through nuanced relationships between abstract concepts such as feelings. While computer vision benchmarks and approaches predominantly focus on understanding and generating literal interpretations of images, metaphorical comprehension of images remains relatively unexplored. Towards this goal, we introduce MetaCLUE, a set of vision tasks on visual metaphor. We also collect high-quality and rich metaphor annotations (abstract objects, concepts, relationships along with their corresponding object boxes) as there do not exist any datasets that facilitate the evaluation of these tasks. We perform a comprehensive analysis of state-of-the-art models in vision and language based on our annotations, highlighting strengths and weaknesses of current approaches in visual metaphor Classification, Localization, Understanding (retrieval, question answering, captioning) and gEneration (text-to-image synthesis) tasks. We hope this work provides a concrete step towards systematically developing AI systems with human-like creative capabilities. Project page: https://metaclue.github.io

****
Towards End-to-End Generative Modeling of Long Videos With Memory-Efficient Bidirectional Transformers

Jaehoon Yoo, Semin Kim, Doyup Lee, Chiheon Kim, Seunghoon Hong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22888-22897

Autoregressive transformers have shown remarkable success in video generation. However, the transformers are prohibited from directly learning the long-term dependency in videos due to the quadratic complexity of self-attention, and inherently suffering from slow inference time and error propagation due to the autoregressive process. In this paper, we propose Memory-efficient Bidirectional Transformer (MeBT) for end-to-end learning of long-term dependency in videos and fast inference. Based on recent advances in bidirectional transformers, our method learns to decode the entire spatio-temporal volume of a video in parallel from partially observed patches. The proposed transformer achieves a linear time complexity in both encoding and decoding, by projecting observable context tokens into a fixed number of latent tokens and conditioning them to decode the masked tokens through the cross-attention. Empowered by linear complexity and bidirectional modeling, our method demonstrates significant improvement over the autoregressive Transformers for generating moderately long videos in both quality and speed.

****
Domain Expansion of Image Generators

Yotam Nitzan, Michaël Gharbi, Richard Zhang, Taesung Park, Jun-Yan Zhu, Daniel Cohen-Or, Eli Shechtman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15933-15942

Can one inject new concepts into an already trained generative model, while respecting its existing structure and knowledge? We propose a new task -- domain expansion -- to address this. Given a pretrained generator and novel (but related) domains, we expand the generator to jointly model all domains, old and new, harmoniously. First, we note the generator contains a meaningful, pretrained latent space. Is it possible to minimally perturb this hard-earned representation, while maximally representing the new domains? Interestingly, we find that the latent space offers unused, "dormant" axes, which do not affect the output. This provides an opportunity -- by "repurposing" these axes, we are able to represent new domains, without perturbing the original representation. In fact, we find that pretrained generators have the capacity to add several -- even hundreds -- of new domains! Using our expansion technique, one "expanded" model can supersede numerous domain-specific models, without expanding model size. Additionally, using a

single, expanded generator natively supports smooth transitions between and com position of domains.
****
On the Effectiveness of Partial Variance Reduction in Federated Learning With He terogeneous Data
Bo Li, Mikkel N. Schmidt, Tommy S. Alstrøm, Sebastian U. Stich; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3964-3973
Data heterogeneity across clients is a key challenge in federated learning. Prio r works address this by either aligning client and server models or using contro l variates to correct client model drift. Although these methods achieve fast co nvergence in convex or simple non-convex problems, the performance in over-param eterized models such as deep neural networks is lacking. In this paper, we first revisit the widely used FedAvg algorithm in a deep neural network to understand how data heterogeneity influences the gradient updates across the neural networ k layers. We observe that while the feature extraction layers are learned effici ently by FedAvg, the substantial diversity of the final classification layers ac ross clients impedes the performance. Motivated by this, we propose to correct m odel drift by variance reduction only on the final layers. We demonstrate that t his significantly outperforms existing benchmarks at a similar or lower communic ation cost. We furthermore provide proof for the convergence rate of our algorit hm.
****
Point Cloud Forecasting as a Proxy for 4D Occupancy Forecasting
Tarasha Khurana, Peiyun Hu, David Held, Deva Ramanan; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1116-1 124
Predicting how the world can evolve in the future is crucial for motion planning in autonomous systems. Classical methods are limited because they rely on costl y human annotations in the form of semantic class labels, bounding boxes, and tr acks or HD maps of cities to plan their motion -- and thus are difficult to scal e to large unlabeled datasets. One promising self-supervised task is 3D point cl oud forecasting from unannotated LiDAR sequences. We show that this task require s algorithms to implicitly capture (1) sensor extrinsics (i.e., the egomotion of the autonomous vehicle), (2) sensor intrinsics (i.e., the sampling pattern spec ific to the particular LiDAR sensor), and (3) the shape and motion of other obje cts in the scene. But autonomous systems should make predictions about the world and not their sensors! To this end, we factor out (1) and (2) by recasting the task as one of spacetime (4D) occupancy forecasting. But because it is expensive to obtain ground-truth 4D occupancy, we "render" point cloud data from 4D occup ancy predictions given sensor extrinsics and intrinsics, allowing one to train a nd test occupancy algorithms with unannotated LiDAR sequences. This also allows one to evaluate and compare point cloud forecasting algorithms across diverse da tasets, sensors, and vehicles.
****
Masked Representation Learning for Domain Generalized Stereo Matching
Zhibo Rao, Bangshu Xiong, Mingyi He, Yuchao Dai, Renjie He, Zhelun Shen, Xing Li ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit ion (CVPR), 2023, pp. 5435-5444
Recently, many deep stereo matching methods have begun to focus on cross-domain performance, achieving impressive achievements. However, these methods did not d eal with the significant volatility of generalization performance among differen t training epochs. Inspired by masked representation learning and multi-task lea rning, this paper designs a simple and effective masked representation for domai n generalized stereo matching. First, we feed the masked left and complete right images as input into the models. Then, we add a lightweight and simple decoder following the feature extraction module to recover the original left image. Fina lly, we train the models with two tasks (stereo matching and image reconstructio n) as a pseudo-multi-task learning framework, promoting models to learn structur e information and to improve generalization performance. We implement our method

on two well-known architectures (CFNet and LacGwcNet) to demonstrate its effectiveness. Experimental results on multi-datasets show that: (1) our method can be easily plugged into the current various stereo matching models to improve generalization performance; (2) our method can reduce the significant volatility of generalization performance among different training epochs; (3) we find that the current methods prefer to choose the best results among different training epochs as generalization performance, but it is impossible to select the best performance by ground truth in practice.
****
LVQAC: Lattice Vector Quantization Coupled With Spatially Adaptive Companding for Efficient Learned Image Compression
Xi Zhang, Xiaolin Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10239-10248
Recently, numerous end-to-end optimized image compression neural networks have been developed and proved themselves as leaders in rate-distortion performance. The main strength of these learnt compression methods is in powerful nonlinear analysis and synthesis transforms that can be facilitated by deep neural networks. However, out of operational expediency, most of these end-to-end methods adopt uniform scalar quantizers rather than vector quantizers, which are information-theoretically optimal. In this paper, we present a novel Lattice Vector Quantization scheme coupled with a spatially Adaptive Companding (LVQAC) mapping. LVQ can better exploit the inter-feature dependencies than scalar uniform quantization while being computationally almost as simple as the latter. Moreover, to improve the adaptability of LVQ to source statistics, we couple a spatially adaptive companding (AC) mapping with LVQ. The resulting LVQAC design can be easily embedded into any end-to-end optimized image compression system. Extensive experiments demonstrate that for any end-to-end CNN image compression models, replacing uniform quantizer by LVQAC achieves better rate-distortion performance without significantly increasing the model complexity.
****
You Can Ground Earlier Than See: An Effective and Efficient Pipeline for Temporal Sentence Grounding in Compressed Videos
Xiang Fang, Daizong Liu, Pan Zhou, Guoshun Nan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2448-2460
Given an untrimmed video, temporal sentence grounding (TSG) aims to locate a target moment semantically according to a sentence query. Although previous respectable works have made decent success, they only focus on high-level visual features extracted from the consecutive decoded frames and fail to handle the compressed videos for query modelling, suffering from insufficient representation capability and significant computational complexity during training and testing. In this paper, we pose a new setting, compressed-domain TSG, which directly utilizes compressed videos rather than fully-decompressed frames as the visual input. To handle the raw video bit-stream input, we propose a novel Three-branch Compressed-domain Spatial-temporal Fusion (TCSF) framework, which extracts and aggregates three kinds of low-level visual features (I-frame, motion vector and residual features) for effective and efficient grounding. Particularly, instead of encoding the whole decoded frames like previous works, we capture the appearance representation by only learning the I-frame feature to reduce delay or latency. Besides, we explore the motion information not only by learning the motion vector feature, but also by exploring the relations of neighboring frames via the residual feature. In this way, a three-branch spatial-temporal attention layer with an adaptive motion-appearance fusion module is further designed to extract and aggregate both appearance and motion information for the final grounding. Experiments on three challenging datasets shows that our TCSF achieves better performance than other state-of-the-art methods with lower complexity.
****
EqMotion: Equivariant Multi-Agent Motion Prediction With Invariant Interaction Reasoning
Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Yu Guang Wang, Xinchao Wang, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt

ern Recognition (CVPR), 2023, pp. 1410-1420
Learning to predict agent motions with relationship reasoning is important for many applications. In motion prediction tasks, maintaining motion equivariance under Euclidean geometric transformations and invariance of agent interaction is a critical and fundamental principle. However, such equivariance and invariance properties are overlooked by most existing methods. To fill this gap, we propose EqMotion, an efficient equivariant motion prediction model with invariant interaction reasoning. To achieve motion equivariance, we propose an equivariant geometric feature learning module to learn a Euclidean transformable feature through dedicated designs of equivariant operations. To reason agent's interactions, we propose an invariant interaction reasoning module to achieve a more stable interaction modeling. To further promote more comprehensive motion features, we propose an invariant pattern feature learning module to learn an invariant pattern feature, which cooperates with the equivariant geometric feature to enhance network expressiveness. We conduct experiments for the proposed model on four distinct scenarios: particle dynamics, molecule dynamics, human skeleton motion prediction and pedestrian trajectory prediction. Experimental results show that our method is not only generally applicable, but also achieves state-of-the-art prediction performances on all the four tasks, improving by 24.0/30.1/8.6/9.2%. Code is available at https://github.com/MediaBrain-SJTU/EqMotion.
****
Fine-Grained Face Swapping via Regional GAN Inversion
Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, Yongwei Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8578-8587
We present a novel paradigm for high-fidelity face swapping that faithfully preserves the desired subtle geometry and texture details. We rethink face swapping from the perspective of fine-grained face editing, i.e., editing for swapping (E4S), and propose a framework that is based on the explicit disentanglement of the shape and texture of facial components. Following the E4S principle, our framework enables both global and local swapping of facial features, as well as controlling the amount of partial swapping specified by the user. Furthermore, the E4S paradigm is inherently capable of handling facial occlusions by means of facial masks. At the core of our system lies a novel Regional GAN Inversion (RGI) method, which allows the explicit disentanglement of shape and texture. It also allows face swapping to be performed in the latent space of StyleGAN. Specifically, we design a multi-scale mask-guided encoder to project the texture of each facial component into regional style codes. We also design a mask-guided injection module to manipulate the feature maps with the style codes. Based on the disentanglement, face swapping is reformulated as a simplified problem of style and mask swapping. Extensive experiments and comparisons with current state-of-the-art methods demonstrate the superiority of our approach in preserving texture and shape details, as well as working with high resolution images. The project page is https://e4s2022.github.io
****
Taming Diffusion Models for Audio-Driven Co-Speech Gesture Generation
Lingting Zhu, Xian Liu, Xuanyu Liu, Rui Qian, Ziwei Liu, Lequan Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10544-10553
Animating virtual avatars to make co-speech gestures facilitates various applications in human-machine interaction. The existing methods mainly rely on generative adversarial networks (GANs), which typically suffer from notorious mode collapse and unstable training, thus making it difficult to learn accurate audio-gesture joint distributions. In this work, we propose a novel diffusion-based framework, named Diffusion Co-Speech Gesture (DiffGesture), to effectively capture the cross-modal audio-to-gesture associations and preserve temporal coherence for high-fidelity audio-driven co-speech gesture generation. Specifically, we first establish the diffusion-conditional generation process on clips of skeleton sequences and audio to enable the whole framework. Then, a novel Diffusion Audio-Gesture Transformer is devised to better attend to the information from multiple mod

alities and model the long-term temporal dependency. Moreover, to eliminate temporal inconsistency, we propose an effective Diffusion Gesture Stabilizer with an annealed noise sampling strategy. Benefiting from the architectural advantages of diffusion models, we further incorporate implicit classifier-free guidance to trade off between diversity and gesture quality. Extensive experiments demonstrate that DiffGesture achieves state-of-the-art performance, which renders coherent gestures with better mode coverage and stronger audio correlations. Code is available at https://github.com/Advocate99/DiffGesture.
****
FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation
Xiaoyu Shi, Zhaoyang Huang, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1599-1610
FlowFormer introduces a transformer architecture into optical flow estimation and achieves state-of-the-art performance. The core component of FlowFormer is the transformer-based cost-volume encoder. Inspired by recent success of masked autoencoding (MAE) pretraining in unleashing transformers' capacity of encoding visual representation, we propose Masked Cost Volume Autoencoding (MCVA) to enhance FlowFormer by pretraining the cost-volume encoder with a novel MAE scheme. Firstly, we introduce a block-sharing masking strategy to prevent masked information leakage, as the cost maps of neighboring source pixels are highly correlated. Secondly, we propose a novel pre-text reconstruction task, which encourages the cost-volume encoder to aggregate long-range information and ensures pretraining-finetuning consistency. We also show how to modify the FlowFormer architecture to accommodate masks during pretraining. Pretrained with MCVA, our proposed FlowFormer++ ranks 1st among published methods on both Sintel and KITTI-2015 benchmarks. Specifically, FlowFormer++ achieves 1.07 and 1.94 average end-point-error (AEPE) on the clean and final pass of Sintel benchmark, leading to 7.76% and 7.18% error reductions from FlowFormer. FlowFormer++ obtains 4.52 F1-all on the KITTI-2015 test set, improving FlowFormer by 0.16.
****
NeRFLix: High-Quality Neural View Synthesis by Learning a Degradation-Driven Inter-Viewpoint MiXer
Kun Zhou, Wenbo Li, Yi Wang, Tao Hu, Nianjuan Jiang, Xiaoguang Han, Jiangbo Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12363-12374
Neural radiance fields(NeRF) show great success in novel-view synthesis. However, in real-world scenes, recovering high-quality details from the source images is still challenging for the existing NeRF-based approaches, due to the potential imperfect calibration information and scene representation inaccuracy. Even with high-quality training frames, the synthetic novel-view frames produced by NeRF models still suffer from notable rendering artifacts, such as noise, blur, etc. Towards to improve the synthesis quality of NeRF-based approaches, we propose NeRFLiX, a general NeRF-agnostic restorer paradigm by learning a degradation-driven inter-viewpoint mixer. Specially, we design a NeRF-style degradation modeling approach and construct large-scale training data, enabling the possibility of effectively removing those NeRF-native rendering artifacts for existing deep neural networks. Moreover, beyond the degradation removal, we propose an inter-viewpoint aggregation framework that is able to fuse highly related high-quality training images, pushing the performance of cutting-edge NeRF models to entirely new levels and producing highly photo-realistic synthetic images.
****
HaLP: Hallucinating Latent Positives for Skeleton-Based Self-Supervised Learning of Actions
Anshul Shah, Aniket Roy, Ketul Shah, Shlok Mishra, David Jacobs, Anoop Cherian, Rama Chellappa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18846-18856
Supervised learning of skeleton sequence encoders for action recognition has received significant attention in recent times. However, learning such encoders wit

hout labels continues to be a challenging problem. While prior works have shown promising results by applying contrastive learning to pose sequences, the qualit y of the learned representations is often observed to be closely tied to data au gmentations that are used to craft the positives. However, augmenting pose seque nces is a difficult task as the geometric constraints among the skeleton joints need to be enforced to make the augmentations realistic for that action. In this work, we propose a new contrastive learning approach to train models for skelet on-based action recognition without labels. Our key contribution is a simple mod ule, HaLP - to Hallucinate Latent Positives for contrastive learning. Specifical ly, HaLP explores the latent space of poses in suitable directions to generate n ew positives. To this end, we present a novel optimization formulation to solve for the synthetic positives with an explicit control on their hardness. We propo se approximations to the objective, making them solvable in closed form with min imal overhead. We show via experiments that using these generated positives with in a standard contrastive learning framework leads to consistent improvements ac ross benchmarks such as NTU-60, NTU-120, and PKU-II on tasks like linear evaluat ion, transfer learning, and kNN evaluation. Our code can be found at https://git hub.com/anshulbshah/HaLP.
****
STMixer: A One-Stage Sparse Action Detector
Tao Wu, Mengqi Cao, Ziteng Gao, Gangshan Wu, Limin Wang; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 147 20-14729
Traditional video action detectors typically adopt the two-stage pipeline, where a person detector is first employed to yield actor boxes and then 3D RoIAlign i s used to extract actor-specific features for classification. This detection par adigm requires multi-stage training and inference and cannot capture context inf ormation outside the bounding box. Recently, a few query-based action detectors are proposed to predict action instances in an end-to-end manner. However, they still lack adaptability in feature sampling or decoding, thus suffering from the issue of inferior performance or slower convergence. In this paper, we propose a new one-stage sparse action detector, termed STMixer. STMixer is based on two core designs. First, we present a query-based adaptive feature sampling module, which endows our STMixer with the flexibility of mining a set of discriminative features from the entire spatiotemporal domain. Second, we devise a dual-branch feature mixing module, which allows our STMixer to dynamically attend to and mix video features along the spatial and the temporal dimension respectively for be tter feature decoding. Coupling these two designs with a video backbone yields a n efficient and accurate action detector. Without bells and whistles, STMixer ob tains the state-of-the-art results on the datasets of AVA, UCF101-24, and JHMDB.
****
3D Human Keypoints Estimation From Point Clouds in the Wild Without Human Labels
Zhenzhen Weng, Alexander S. Gorban, Jingwei Ji, Mahyar Najibi, Yin Zhou, Dragomi r Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2023, pp. 1158-1167
Training a 3D human keypoint detector from point clouds in a supervised manner r equires large volumes of high quality labels. While it is relatively easy to cap ture large amounts of human point clouds, annotating 3D keypoints is expensive, subjective, error prone and especially difficult for long-tail cases (pedestrian s with rare poses, scooterists, etc.). In this work, we propose GC-KPL - Geometr y Consistency inspired Key Point Leaning, an approach for learning 3D human join t locations from point clouds without human labels. We achieve this by our novel unsupervised loss formulations that account for the structure and movement of t he human body. We show that by training on a large training set from Waymo Open Dataset without any human annotated keypoints, we are able to achieve reasonable performance as compared to the fully supervised approach. Further, the backbone benefits from the unsupervised training and is useful in downstream fewshot lea rning of keypoints, where fine-tuning on only 10 percent of the labeled training data gives comparable performance to fine-tuning on the entire set. We demonstr ated that GC-KPL outperforms by a large margin over SoTA when trained on entire

dataset and efficiently leverages large volumes of unlabeled data.
****
Where Is My Spot? Few-Shot Image Generation via Latent Subspace Optimization
Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3272-3281
Image generation relies on massive training data that can hardly produce diverse images of an unseen category according to a few examples. In this paper, we address this dilemma by projecting sparse few-shot samples into a continuous latent space that can potentially generate infinite unseen samples. The rationale behind is that we aim to locate a centroid latent position in a conditional StyleGAN, where the corresponding output image on that centroid can maximize the similarity with the given samples. Although the given samples are unseen for the conditional StyleGAN, we assume the neighboring latent subspace around the centroid belongs to the novel category, and therefore introduce two latent subspace optimization objectives. In the first one we use few-shot samples as positive anchors of the novel class, and adjust the StyleGAN to produce the corresponding results with the new class label condition. The second objective is to govern the generation process from the other way around, by altering the centroid and its surrounding latent subspace for a more precise generation of the novel class. These reciprocal optimization objectives inject a novel class into the StyleGAN latent subspace, and therefore new unseen samples can be easily produced by sampling images from it. Extensive experiments demonstrate superior few-shot generation performances compared with state-of-the-art methods, especially in terms of diversity and generation quality. Code is available at https://github.com/chansey0529/LSO.
****
FLEX: Full-Body Grasping Without Full-Body Grasps
Purva Tendulkar, Dídac Surís, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21179-21189
Synthesizing 3D human avatars interacting realistically with a scene is an important problem with applications in AR/VR, video games, and robotics. Towards this goal, we address the task of generating a virtual human -- hands and full body -- grasping everyday objects. Existing methods approach this problem by collecting a 3D dataset of humans interacting with objects and training on this data. However, 1) these methods do not generalize to different object positions and orientations or to the presence of furniture in the scene, and 2) the diversity of their generated full-body poses is very limited. In this work, we address all the above challenges to generate realistic, diverse full-body grasps in everyday scenes without requiring any 3D full-body grasping data. Our key insight is to leverage the existence of both full-body pose and hand-grasping priors, composing them using 3D geometrical constraints to obtain full-body grasps. We empirically validate that these constraints can generate a variety of feasible human grasps that are superior to baselines both quantitatively and qualitatively.
****
Genie: Show Me the Data for Quantization
Yongkweon Jeon, Chungman Lee, Ho-young Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12064-12073
Zero-shot quantization is a promising approach for developing lightweight deep neural networks when data is inaccessible owing to various reasons, including cost and issues related to privacy. By exploiting the learned parameters (u and sigma) of batch normalization layers in an FP32-pre-trained model, zero-shot quantization schemes focus on generating synthetic data. Subsequently, they distill knowledge from the pre-trained model (teacher) to the quantized model (student) such that the quantized model can be optimized with the synthetic dataset. However, thus far, zero-shot quantization has primarily been discussed in the context of quantization-aware training methods, which require task-specific losses and long-term optimization as much as retraining. We thus introduce a post-training quantization scheme for zero-shot quantization that produces high-quality quantized networks within a few hours. Furthermore, we propose a framework called GENIE

that generates data suited for quantization. With the data synthesized by GENIE, we can produce robust quantized models without real datasets, which is comparable to few-shot quantization. We also propose a post-training quantization algorithm to enhance the performance of quantized models. By combining them, we can bridge the gap between zero-shot and few-shot quantization while significantly improving the quantization performance compared to that of existing approaches. In other words, we can obtain a unique state-of-the-art zero-shot quantization approach.
****
EVA: Exploring the Limits of Masked Visual Representation Learning at Scale
Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, Yue Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19358-19369
We launch EVA, a vision-centric foundation model to explore the limits of visual representation at scale using only publicly accessible data. EVA is a vanilla ViT pre-trained to reconstruct the masked out image-text aligned vision features conditioned on visible image patches. Via this pretext task, we can efficiently scale up EVA to one billion parameters, and sets new records on a broad range of representative vision downstream tasks, such as image recognition, video action recognition, object detection, instance segmentation and semantic segmentation without heavy supervised training. Moreover, we observe quantitative changes in scaling EVA result in qualitative changes in transfer learning performance that are not present in other models. For instance, EVA takes a great leap in the challenging large vocabulary instance segmentation task: our model achieves almost the same state-of-the-art performance on LVIS dataset with over a thousand categories and COCO dataset with only eighty categories. Beyond a pure vision encoder, EVA can also serve as a vision-centric, multi-modal pivot to connect images and text. We find initializing the vision tower of a giant CLIP from EVA can greatly stabilize the training and outperform the training from scratch counterpart with much fewer samples and less compute, providing a new direction for scaling up and accelerating the costly training of multi-modal foundation models.
****
TopNet: Transformer-Based Object Placement Network for Image Compositing
Sijie Zhu, Zhe Lin, Scott Cohen, Jason Kuen, Zhifei Zhang, Chen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1838-1847
We investigate the problem of automatically placing an object into a background image for image compositing. Given a background image and a segmented object, the goal is to train a model to predict plausible placements (location and scale) of the object for compositing. The quality of the composite image highly depends on the predicted location/scale. Existing works either generate candidate bounding boxes or apply sliding-window search using global representations from background and object images, which fail to model local information in background images. However, local clues in background images are important to determine the compatibility of placing the objects with certain locations/scales. In this paper, we propose to learn the correlation between object features and all local background features with a transformer module so that detailed information can be provided on all possible location/scale configurations. A sparse contrastive loss is further proposed to train our model with sparse supervision. Our new formulation generates a 3D heatmap indicating the plausibility of all location/scale combinations in one network forward pass, which is >10x faster than the previous sliding-window method. It also supports interactive search when users provide a pre-defined location or scale. The proposed method can be trained with explicit annotation or in a self-supervised manner using an off-the-shelf inpainting model, and it outperforms state-of-the-art methods significantly. User study shows that the trained model generalizes well to real-world images with diverse challenging scenes and object categories.
****
Discrete Point-Wise Attack Is Not Enough: Generalized Manifold Adversarial Attack for Face Recognition

Qian Li, Yuxiao Hu, Ye Liu, Dongxiao Zhang, Xin Jin, Yuntian Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20575-20584
Classical adversarial attacks for Face Recognition (FR) models typically generate discrete examples for target identity with a single state image. However, such paradigm of point-wise attack exhibits poor generalization against numerous unknown states of identity and can be easily defended. In this paper, by rethinking the inherent relationship between the face of target identity and its variants, we introduce a new pipeline of Generalized Manifold Adversarial Attack (GMAA) to achieve a better attack performance by expanding the attack range. Specifically, this expansion lies on two aspects -- GMAA not only expands the target to be attacked from one to many to encourage a good generalization ability for the generated adversarial examples, but it also expands the latter from discrete points to manifold by leveraging the domain knowledge that face expression change can be continuous, which enhances the attack effect as a data augmentation mechanism did. Moreover, we further design a dual supervision with local and global constraints as a minor contribution to improve the visual quality of the generated adversarial examples. We demonstrate the effectiveness of our method based on extensive experiments, and reveal that GMAA promises a semantic continuous adversarial space with a higher generalization ability and visual quality.
****
Gloss Attention for Gloss-Free Sign Language Translation
Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, Zhou Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2551-2562
Most sign language translation (SLT) methods to date require the use of gloss annotations to provide additional supervision information, however, the acquisition of gloss is not easy. To solve this problem, we first perform an analysis of existing models to confirm how gloss annotations make SLT easier. We find that it can provide two aspects of information for the model, 1) it can help the model implicitly learn the location of semantic boundaries in continuous sign language videos, 2) it can help the model understand the sign language video globally. We then propose gloss attention, which enables the model to keep its attention within video segments that have the same semantics locally, just as gloss helps existing models do. Furthermore, we transfer the knowledge of sentence-to-sentence similarity from the natural language model to our gloss attention SLT network (GASLT) to help it understand sign language videos at the sentence level. Experimental results on multiple large-scale sign language datasets show that our proposed GASLT model significantly outperforms existing methods. Our code is provided in https://github.com/YinAoXiong/GASLT.
****
Multi-Agent Automated Machine Learning
Zhaozhi Wang, Kefan Su, Jian Zhang, Huizhu Jia, Qixiang Ye, Xiaodong Xie, Zongqing Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11960-11969
In this paper, we propose multi-agent automated machine learning (MA2ML) with the aim to effectively handle joint optimization of modules in automated machine learning (AutoML). MA2ML takes each machine learning module, such as data augmentation (AUG), neural architecture search (NAS), or hyper-parameters (HPO), as an agent and the final performance as the reward, to formulate a multi-agent reinforcement learning problem. MA2ML explicitly assigns credit to each agent according to its marginal contribution to enhance cooperation among modules, and incorporates off-policy learning to improve search efficiency. Theoretically, MA2ML guarantees monotonic improvement of joint optimization. Extensive experiments show that MA2ML yields the state-of-the-art top-1 accuracy on ImageNet under constraints of computational cost, e.g., 79.7%/80.5% with FLOPs fewer than 600M/800M. Extensive ablation studies verify the benefits of credit assignment and off-policy learning of MA2ML.
****
Robot Structure Prior Guided Temporal Attention for Camera-to-Robot Pose Estimat

ion From Image Sequence
Yang Tian, Jiyao Zhang, Zekai Yin, Hao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8917-8926
In this work, we tackle the problem of online camera-to-robot pose estimation from single-view successive frames of an image sequence, a crucial task for robots to interact with the world. The primary obstacles of this task are the robot's self-occlusions and the ambiguity of single-view images. This work demonstrates, for the first time, the effectiveness of temporal information and the robot structure prior in addressing these challenges. Given the successive frames and the robot joint configuration, our method learns to accurately regress the 2D coordinates of the predefined robot's keypoints (e.g., joints). With the camera intrinsic and robotic joints status known, we get the camera-to-robot pose using a Perspective-n-point (PnP) solver. We further improve the camera-to-robot pose iteratively using the robot structure prior. To train the whole pipeline, we build a large-scale synthetic dataset generated with domain randomisation to bridge the sim-to-real gap. The extensive experiments on synthetic and real-world datasets and the downstream robotic grasping task demonstrate that our method achieves new state-of-the-art performances and outperforms traditional hand-eye calibration algorithms in real-time (36 FPS). Code and data are available at the project page: https://sites.google.com/view/sgtapose.
****
FREDOM: Fairness Domain Adaptation Approach to Semantic Scene Understanding
Thanh-Dat Truong, Ngan Le, Bhiksha Raj, Jackson Cothren, Khoa Luu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19988-19997
Although Domain Adaptation in Semantic Scene Segmentation has shown impressive improvement in recent years, the fairness concerns in the domain adaptation have yet to be well defined and addressed. In addition, fairness is one of the most critical aspects when deploying the segmentation models into human-related real-world applications, e.g., autonomous driving, as any unfair predictions could influence human safety. In this paper, we propose a novel Fairness Domain Adaptation (FREDOM) approach to semantic scene segmentation. In particular, from the proposed formulated fairness objective, a new adaptation framework will be introduced based on the fair treatment of class distributions. Moreover, to generally model the context of structural dependency, a new conditional structural constraint is introduced to impose the consistency of predicted segmentation. Thanks to the proposed Conditional Structure Network, the self-attention mechanism has sufficiently modeled the structural information of segmentation. Through the ablation studies, the proposed method has shown the performance improvement of the segmentation models and promoted fairness in the model predictions. The experimental results on the two standard benchmarks, i.e., SYNTHIA -> Cityscapes and GTA5 -> Cityscapes, have shown that our method achieved State-of-the-Art (SOTA) performance.
****
IMP: Iterative Matching and Pose Estimation With Adaptive Pooling
Fei Xue, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21317-21326
Previous methods solve feature matching and pose estimation using a two-stage process by first finding matches and then estimating the pose. As they ignore the geometric relationships between the two tasks, they focus on either improving the quality of matches or filtering potential outliers, leading to limited efficiency or accuracy. In contrast, we propose an iterative matching and pose estimation framework (IMP) leveraging the geometric connections between the two tasks: a few good matches are enough for a roughly accurate pose estimation; a roughly accurate pose can be used to guide the matching by providing geometric constraints. To this end, we implement a geometry-aware recurrent module with transformers which jointly outputs sparse matches and camera poses. Specifically, for each iteration, we first implicitly embed geometric information into the module via a pose-consistency loss, allowing it to predict geometry-aware matches progressively. Second, we introduce an efficient IMP (EIMP) to dynamically discard keypoint

s without potential matches, avoiding redundant updating and significantly reducing the quadratic time complexity of attention computation in transformers. Experiments on YFCC100m, Scannet, and Aachen Day-Night datasets demonstrate that the proposed method outperforms previous approaches in terms of accuracy and efficiency.
****
HRDFuse: Monocular 360deg Depth Estimation by Collaboratively Learning Holistic-With-Regional Depth Distributions
Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13273-13282
Depth estimation from a monocular 360 image is a burgeoning problem owing to its holistic sensing of a scene. Recently, some methods, e.g., OmniFusion, have applied the tangent projection (TP) to represent a 360 image and predicted depth values via patch-wise regressions, which are merged to get a depth map with equirectangular projection (ERP) format. However, these methods suffer from 1) non-trivial process of merging a large number of patches; 2) capturing less holistic-with-regional contextual information by directly regressing the depth value of each pixel. In this paper, we propose a novel framework, HRDFuse, that subtly combines the potential of convolutional neural networks (CNNs) and transformers by collaboratively learning the holistic contextual information from the ERP and the regional structural information from the TP. Firstly, we propose a spatial feature alignment (SFA) module that learns feature similarities between the TP and ERP to aggregate the TP features into a complete ERP feature map in a pixel-wise manner. Secondly, we propose a collaborative depth distribution classification (CDDC) module that learns the holistic-with-regional histograms capturing the ERP and TP depth distributions. As such, the final depth values can be predicted as a linear combination of histogram bin centers. Lastly, we adaptively combine the depth predictions from ERP and TP to obtain the final depth map. Extensive experiments show that our method predicts more smooth and accurate depth results while achieving favorably better results than the SOTA methods.
****
Revisiting Rolling Shutter Bundle Adjustment: Toward Accurate and Fast Solution
Bangyan Liao, Delin Qu, Yifei Xue, Huiqing Zhang, Yizhen Lao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4863-4871
We propose a robust and fast bundle adjustment solution that estimates the 6-DoF pose of the camera and the geometry of the environment based on measurements from a rolling shutter (RS) camera. This tackles the challenges in the existing works, namely relying on additional sensors, high frame rate video as input, restrictive assumptions on camera motion, readout direction, and poor efficiency. To this end, we first investigate the influence of normalization to the image point on RSBA performance and show its better approximation in modelling the real 6-DoF camera motion. Then we present a novel analytical model for the visual residual covariance, which can be used to standardize the reprojection error during the optimization, consequently improving the overall accuracy. More importantly, the combination of normalization and covariance standardization weighting in RSBA (NW-RSBA) can avoid common planar degeneracy without needing to constrain the filming manner. Besides, we propose an acceleration strategy for NW-RSBA based on the sparsity of its Jacobian matrix and Schur complement. The extensive synthetic and real data experiments verify the effectiveness and efficiency of the proposed solution over the state-of-the-art works. We also demonstrate the proposed method can be easily implemented and plug-in famous GSSfM and GSSLAM systems as completed RSSfM and RSSLAM solutions.
****
StructVPR: Distill Structural Knowledge With Weighting Samples for Visual Place Recognition
Yanqing Shen, Sanping Zhou, Jingwen Fu, Ruotong Wang, Shitao Chen, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11217-11226

Visual place recognition (VPR) is usually considered as a specific image retrieval problem. Limited by existing training frameworks, most deep learning-based works cannot extract sufficiently stable global features from RGB images and rely on a time-consuming re-ranking step to exploit spatial structural information for better performance. In this paper, we propose StructVPR, a novel training architecture for VPR, to enhance structural knowledge in RGB global features and thus improve feature stability in a constantly changing environment. Specifically, StructVPR uses segmentation images as a more definitive source of structural knowledge input into a CNN network and applies knowledge distillation to avoid online segmentation and inference of seg-branch in testing. Considering that not all samples contain high-quality and helpful knowledge, and some even hurt the performance of distillation, we partition samples and weigh each sample's distillation loss to enhance the expected knowledge precisely. Finally, StructVPR achieves impressive performance on several benchmarks using only global retrieval and even outperforms many two-stage approaches by a large margin. After adding additional re-ranking, ours achieves state-of-the-art performance while maintaining a low computational cost.
****

PATS: Patch Area Transportation With Subdivision for Local Feature Matching
Junjie Ni, Yijin Li, Zhaoyang Huang, Hongsheng Li, Hujun Bao, Zhaopeng Cui, Guofeng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17776-17786
Local feature matching aims at establishing sparse correspondences between a pair of images. Recently, detector-free methods present generally better performance but are not satisfactory in image pairs with large scale differences. In this paper, we propose Patch Area Transportation with Subdivision (PATS) to tackle this issue. Instead of building an expensive image pyramid, we start by splitting the original image pair into equal-sized patches and gradually resizing and subdividing them into smaller patches with the same scale. However, estimating scale differences between these patches is non-trivial since the scale differences are determined by both relative camera poses and scene structures, and thus spatially varying over image pairs. Moreover, it is hard to obtain the ground truth for real scenes. To this end, we propose patch area transportation, which enables learning scale differences in a self-supervised manner. In contrast to bipartite graph matching, which only handles one-to-one matching, our patch area transportation can deal with many-to-many relationships. PATS improves both matching accuracy and coverage, and shows superior performance in downstream tasks, such as relative pose estimation, visual localization, and optical flow estimation.The source code will be released to benefit the community.
****

Learning Human-to-Robot Handovers From Point Clouds
Sammy Christen, Wei Yang, Claudia Pérez-D’Arpino, Otmar Hilliges, Dieter Fox, Yu-Wei Chao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9654-9664
We propose the first framework to learn control policies for vision-based human-to-robot handovers, a critical task for human-robot interaction. While research in Embodied AI has made significant progress in training robot agents in simulated environments, interacting with humans remains challenging due to the difficulties of simulating humans. Fortunately, recent research has developed realistic simulated environments for human-to-robot handovers. Leveraging this result, we introduce a method that is trained with a human-in-the-loop via a two-stage teacher-student framework that uses motion and grasp planning, reinforcement learning, and self-supervision. We show significant performance gains over baselines on a simulation benchmark, sim-to-sim transfer and sim-to-real transfer.
****

MEDIC: Remove Model Backdoors via Importance Driven Cloning
Qiuling Xu, Guanhong Tao, Jean Honorio, Yingqi Liu, Shengwei An, Guangyu Shen, Siyuan Cheng, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20485-20494
We develop a novel method to remove injected backdoors in deep learning models.

It works by cloning the benign behaviors of a trojaned model to a new model of the same structure. It trains the clone model from scratch on a very small subset of samples and aims to minimize a cloning loss that denotes the differences between the activations of important neurons across the two models. The set of important neurons varies for each input, depending on their magnitude of activations and their impact on the classification result. We theoretically show our method can better recover benign functions of the backdoor model. Meanwhile, we prove our method can be more effective in removing backdoors compared with fine-tuning. Our experiments show that our technique can effectively remove nine different types of backdoors with minor benign accuracy degradation, outperforming the state-of-the-art backdoor removal techniques that are based on fine-tuning, knowledge distillation, and neuron pruning.
****
Context-Aware Relative Object Queries To Unify Video Instance and Panoptic Segmentation
Anwesa Choudhuri, Girish Chowdhary, Alexander G. Schwing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6377-6386
Object queries have emerged as a powerful abstraction to generically represent object proposals. However, their use for temporal tasks like video segmentation poses two questions: 1) How to process frames sequentially and propagate object queries seamlessly across frames. Using independent object queries per frame doesn't permit tracking, and requires post-processing. 2) How to produce temporally consistent, yet expressive object queries that model both appearance and position changes. Using the entire video at once doesn't capture position changes and doesn't scale to long videos. As one answer to both questions we propose 'context-aware relative object queries', which are continuously propagated frame-by-frame. They seamlessly track objects and deal with occlusion and re-appearance of objects, without post-processing. Further, we find context-aware relative object queries better capture position changes of objects in motion. We evaluate the proposed approach across three challenging tasks: video instance segmentation, multi-object tracking and segmentation, and video panoptic segmentation. Using the same approach and architecture, we match or surpass state-of-the art results on the diverse and challenging OVIS, Youtube-VIS, Cityscapes-VPS, MOTS 2020 and KITTI-MOTS data.
****
Score Jacobian Chaining: Lifting Pretrained 2D Diffusion Models for 3D Generation
Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, Greg Shakhnarovich; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12619-12629
A diffusion model learns to predict a vector field of gradients. We propose to apply chain rule on the learned gradients, and back-propagate the score of a diffusion model through the Jacobian of a differentiable renderer, which we instantiate to be a voxel radiance field. This setup aggregates 2D scores at multiple camera viewpoints into a 3D score, and repurposes a pretrained 2D model for 3D data generation. We identify a technical challenge of distribution mismatch that arises in this application, and propose a novel estimation mechanism to resolve it. We run our algorithm on several off-the-shelf diffusion image generative models, including the recently released Stable Diffusion trained on the large-scale LAION dataset.
****
Role of Transients in Two-Bounce Non-Line-of-Sight Imaging
Siddharth Somasundaram, Akshat Dave, Connor Henley, Ashok Veeraraghavan, Ramesh Raskar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9192-9201
The goal of non-line-of-sight (NLOS) imaging is to image objects occluded from the camera's field of view using multiply scattered light. Recent works have demonstrated the feasibility of two-bounce (2B) NLOS imaging by scanning a laser and measuring cast shadows of occluded objects in scenes with two relay surfaces. I

n this work, we study the role of time-of-flight (ToF) measurements, i.e. transi
ents, in 2B-NLOS under multiplexed illumination. Specifically, we study how ToF
information can reduce the number of measurements and spatial resolution needed
for shape reconstruction. We present our findings with respect to tradeoffs in (
1) temporal resolution, (2) spatial resolution, and (3) number of image captures
 by studying SNR and recoverability as functions of system parameters. This lead
s to a formal definition of the mathematical constraints for 2B lidar. We believ
e that our work lays an analytical groundwork for design of future NLOS imaging
systems, especially as ToF sensors become increasingly ubiquitous.
****
SimpleNet: A Simple Network for Image Anomaly Detection and Localization
Zhikang Liu, Yiming Zhou, Yuansheng Xu, Zilei Wang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20402-20
411
We propose a simple and application-friendly network (called SimpleNet) for dete
cting and localizing anomalies. SimpleNet consists of four components: (1) a pre
-trained Feature Extractor that generates local features, (2) a shallow Feature
Adapter that transfers local features towards target domain, (3) a simple Anomal
y Feature Generator that counterfeits anomaly features by adding Gaussian noise
to normal features, and (4) a binary Anomaly Discriminator that distinguishes an
omaly features from normal features. During inference, the Anomaly Feature Gener
ator would be discarded. Our approach is based on three intuitions. First, trans
forming pre-trained features to target-oriented features helps avoid domain bias
. Second, generating synthetic anomalies in feature space is more effective, as
defects may not have much commonality in the image space. Third, a simple discri
minator is much efficient and practical. In spite of simplicity, SimpleNet outpe
rforms previous methods quantitatively and qualitatively. On the MVTec AD benchm
ark, SimpleNet achieves an anomaly detection AUROC of 99.6%, reducing the error
by 55.5% compared to the next best performing model. Furthermore, SimpleNet is f
aster than existing methods, with a high frame rate of 77 FPS on a 3080ti GPU. A
dditionally, SimpleNet demonstrates significant improvements in performance on t
he One-Class Novelty Detection task. Code: https://github.com/DonaldRR/SimpleNet
.
****
Elastic Aggregation for Federated Optimization
Dengsheng Chen, Jie Hu, Vince Junkai Tan, Xiaoming Wei, Enhua Wu; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 12187-12197
Federated learning enables the privacy-preserving training of neural network mod
els using real-world data across distributed clients. FedAvg has become the pref
erred optimizer for federated learning because of its simplicity and effectivene
ss. FedAvg uses naive aggregation to update the server model, interpolating clie
nt models based on the number of instances used in their training. However, naiv
e aggregation suffers from client-drift when the data is heterogenous (non-IID),
 leading to unstable and slow convergence. In this work, we propose a novel aggr
egation approach, elastic aggregation, to overcome these issues. Elastic aggrega
tion interpolates client models adaptively according to parameter sensitivity, w
hich is measured by computing how much the overall prediction function output ch
anges when each parameter is changed. This measurement is performed in an unsupe
rvised and online manner. Elastic aggregation reduces the magnitudes of updates
to the more sensitive parameters so as to prevent the server model from drifting
 to any one client distribution, and conversely boosts updates to the less sensi
tive parameters to better explore different client distributions. Empirical resu
lts on real and synthetic data as well as analytical results show that elastic a
ggregation leads to efficient training in both convex and non-convex settings, w
hile being fully agnostic to client heterogeneity and robust to large numbers of
 clients, partial participation, and imbalanced data. Finally, elastic aggregati
on works well with other federated optimizers and achieves significant improveme
nts across the board.
****

G-MSM: Unsupervised Multi-Shape Matching With Graph-Based Affinity Priors
Marvin Eisenberger, Aysim Toker, Laura Leal-Taixé, Daniel Cremers; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22762-22772
We present G-MSM (Graph-based Multi-Shape Matching), a novel unsupervised learning approach for non-rigid shape correspondence. Rather than treating a collection of input poses as an unordered set of samples, we explicitly model the underlying shape data manifold. To this end, we propose an adaptive multi-shape matching architecture that constructs an affinity graph on a given set of training shapes in a self-supervised manner. The key idea is to combine putative, pairwise correspondences by propagating maps along shortest paths in the underlying shape graph. During training, we enforce cycle-consistency between such optimal paths and the pairwise matches which enables our model to learn topology-aware shape priors. We explore different classes of shape graphs and recover specific settings, like template-based matching (star graph) or learnable ranking/sorting (TSP graph), as special cases in our framework. Finally, we demonstrate state-of-the-art performance on several recent shape correspondence benchmarks, including real-world 3D scan meshes with topological noise and challenging inter-class pairs.
****
Enhancing Deformable Local Features by Jointly Learning To Detect and Describe Keypoints
Guilherme Potje, Felipe Cadar, André Araujo, Renato Martins, Erickson R. Nascimento; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1306-1315
Local feature extraction is a standard approach in computer vision for tackling important tasks such as image matching and retrieval. The core assumption of most methods is that images undergo affine transformations, disregarding more complicated effects such as non-rigid deformations. Furthermore, incipient works tailored for non-rigid correspondence still rely on keypoint detectors designed for rigid transformations, hindering performance due to the limitations of the detector. We propose DALF (Deformation-Aware Local Features), a novel deformation-aware network for jointly detecting and describing keypoints, to handle the challenging problem of matching deformable surfaces. All network components work cooperatively through a feature fusion approach that enforces the descriptors' distinctiveness and invariance. Experiments using real deforming objects showcase the superiority of our method, where it delivers 8% improvement in matching scores compared to the previous best results. Our approach also enhances the performance of two real-world applications: deformable object retrieval and non-rigid 3D surface registration. Code for training, inference, and applications are publicly available at verlab.dcc.ufmg.br/descriptors/dalf_cvpr23.
****
ObjectMatch: Robust Registration Using Canonical Object Correspondences
Can Gümeli, Angela Dai, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13082-13091
We present ObjectMatch, a semantic and object-centric camera pose estimator for RGB-D SLAM pipelines. Modern camera pose estimators rely on direct correspondences of overlapping regions between frames; however, they cannot align camera frames with little or no overlap. In this work, we propose to leverage indirect correspondences obtained via semantic object identification. For instance, when an object is seen from the front in one frame and from the back in another frame, we can provide additional pose constraints through canonical object correspondences. We first propose a neural network to predict such correspondences on a per-pixel level, which we then combine in our energy formulation with state-of-the-art keypoint matching solved with a joint Gauss-Newton optimization. In a pairwise setting, our method improves registration recall of state-of-the-art feature matching, including from 24% to 45% in pairs with 10% or less inter-frame overlap. In registering RGB-D sequences, our method outperforms cutting-edge SLAM baselines in challenging, low-frame-rate scenarios, achieving more than 35% reduction in trajectory error in multiple scenes.
****

Siamese Image Modeling for Self-Supervised Vision Representation Learning
Chenxin Tao, Xizhou Zhu, Weijie Su, Gao Huang, Bin Li, Jie Zhou, Yu Qiao, Xiaogang Wang, Jifeng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2132-2141
Self-supervised learning (SSL) has delivered superior performance on a variety of downstream vision tasks. Two main-stream SSL frameworks have been proposed, i.e., Instance Discrimination (ID) and Masked Image Modeling (MIM). ID pulls together representations from different views of the same image, while avoiding feature collapse. It lacks spatial sensitivity, which requires modeling the local structure within each image. On the other hand, MIM reconstructs the original content given a masked image. It instead does not have good semantic alignment, which requires projecting semantically similar views into nearby representations. To address this dilemma, we observe that (1) semantic alignment can be achieved by matching different image views with strong augmentations; (2) spatial sensitivity can benefit from predicting dense representations with masked images. Driven by these analysis, we propose Siamese Image Modeling (SiameseIM), which predicts the dense representations of an augmented view, based on another masked view from the same image but with different augmentations. SiameseIM uses a Siamese network with two branches. The online branch encodes the first view, and predicts the second view's representation according to the relative positions between these two views. The target branch produces the target by encoding the second view. SiameseIM can surpass both ID and MIM on a wide range of downstream tasks, including ImageNet finetuning and linear probing, COCO and LVIS detection, and ADE20k semantic segmentation. The improvement is more significant in few-shot, long-tail and robustness-concerned scenarios. Code shall be released.
****
Generating Part-Aware Editable 3D Shapes Without 3D Supervision
Konstantinos Tertikas, Despoina Paschalidou, Boxiao Pan, Jeong Joon Park, Mikaela Angelina Uy, Ioannis Emiris, Yannis Avrithis, Leonidas Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4466-4478
Impressive progress in generative models and implicit representations gave rise to methods that can generate 3D shapes of high quality. However, being able to locally control and edit shapes is another essential property that can unlock several content creation applications. Local control can be achieved with part-aware models, but existing methods require 3D supervision and cannot produce textures. In this work, we devise PartNeRF, a novel part-aware generative model for editable 3D shape synthesis that does not require any explicit 3D supervision. Our model generates objects as a set of locally defined NeRFs, augmented with an affine transformation. This enables several editing operations such as applying transformations on parts, mixing parts from different objects etc. To ensure distinct, manipulable parts we enforce a hard assignment of rays to parts that makes sure that the color of each ray is only determined by a single NeRF. As a result, altering one part does not affect the appearance of the others. Evaluations on various ShapeNet categories demonstrate the ability of our model to generate editable 3D objects of improved fidelity, compared to previous part-based generative approaches that require 3D supervision or models relying on NeRFs.
****
Center Focusing Network for Real-Time LiDAR Panoptic Segmentation
Xiaoyan Li, Gang Zhang, Boyue Wang, Yongli Hu, Baocai Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13425-13434
LiDAR panoptic segmentation facilitates an autonomous vehicle to comprehensively understand the surrounding objects and scenes and is required to run in real time. The recent proposal-free methods accelerate the algorithm, but their effectiveness and efficiency are still limited owing to the difficulty of modeling non-existent instance centers and the costly center-based clustering modules. To achieve accurate and real-time LiDAR panoptic segmentation, a novel center focusing network (CFNet) is introduced. Specifically, the center focusing feature encoding (CFFE) is proposed to explicitly understand the relationships between the ori

ginal LiDAR points and virtual instance centers by shifting the LiDAR points and filling in the center points. Moreover, to leverage the redundantly detected centers, a fast center deduplication module (CDM) is proposed to select only one center for each instance. Experiments on the SemanticKITTI and nuScenes panoptic segmentation benchmarks demonstrate that our CFNet outperforms all existing methods by a large margin and is 1.6 times faster than the most efficient method.
****
High-Fidelity Facial Avatar Reconstruction From Monocular Video With Generative Priors
Yunpeng Bai, Yanbo Fan, Xuan Wang, Yong Zhang, Jingxiang Sun, Chun Yuan, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4541-4551
High-fidelity facial avatar reconstruction from a monocular video is a significant research problem in computer graphics and computer vision. Recently, Neural Radiance Field (NeRF) has shown impressive novel view rendering results and has been considered for facial avatar reconstruction. However, the complex facial dynamics and missing 3D information in monocular videos raise significant challenges for faithful facial reconstruction. In this work, we propose a new method for NeRF-based facial avatar reconstruction that utilizes 3D-aware generative prior. Different from existing works that depend on a conditional deformation field for dynamic modeling, we propose to learn a personalized generative prior, which is formulated as a local and low dimensional subspace in the latent space of 3D-GAN. We propose an efficient method to construct the personalized generative prior based on a small set of facial images of a given individual. After learning, it allows for photo-realistic rendering with novel views, and the face reenactment can be realized by performing navigation in the latent space. Our proposed method is applicable for different driven signals, including RGB images, 3DMM coefficients, and audio. Compared with existing works, we obtain superior novel view synthesis results and faithfully face reenactment performance. The code is available here https://github.com/bbaaii/HFA-GP.
****
Mixed Autoencoder for Self-Supervised Visual Representation Learning
Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22742-22751
Masked Autoencoder (MAE) has demonstrated superior performance on various vision tasks via randomly masking image patches and reconstruction. However, effective data augmentation strategies for MAE still remain open questions, different from those in contrastive learning that serve as the most important part. This paper studies the prevailing mixing augmentation for MAE. We first demonstrate that naive mixing will in contrast degenerate model performance due to the increase of mutual information (MI). To address, we propose homologous recognition, an auxiliary pretext task, not only to alleviate the MI increasement by explicitly requiring each patch to recognize homologous patches, but also to perform object-aware self-supervised pre-training for better downstream dense perception performance. With extensive experiments, we demonstrate that our proposed Mixed Autoencoder (MixedAE) achieves the state-of-the-art transfer results among masked image modeling (MIM) augmentations on different downstream tasks with significant efficiency. Specifically, our MixedAE outperforms MAE by +0.3% accuracy, +1.7 mIoU and +0.9 AP on ImageNet-1K, ADE20K and COCO respectively with a standard ViT-Base. Moreover, MixedAE surpasses iBOT, a strong MIM method combined with instance discrimination, while accelerating training by 2x. To our best knowledge, this is the very first work to consider mixing for MIM from the perspective of pretext task design. Code will be made available.
****
Restoration of Hand-Drawn Architectural Drawings Using Latent Space Mapping With Degradation Generator
Nakkwan Choi, Seungjae Lee, Yongsik Lee, Seungjoon Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14164-14172

This work presents the restoration of drawings of wooden built heritage. Hand-drawn drawings contain the most important original information but are often severely degraded over time. A novel restoration method based on the vector quantized variational autoencoders is presented. Latent space representations of drawings and noise are learned, which are used to map noisy drawings to clean drawings for restoration and to generate authentic noisy drawings for data augmentation. The proposed method is applied to the drawings archived in the Cultural Heritage Administration. Restored drawings show significant quality improvement and allow more accurate interpretations of information.
****
CABM: Content-Aware Bit Mapping for Single Image Super-Resolution Network With Large Input
Senmao Tian, Ming Lu, Jiaming Liu, Yandong Guo, Yurong Chen, Shunli Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1756-1765
With the development of high-definition display devices, the practical scenario of Super-Resolution (SR) usually needs to super-resolve large input like 2K to higher resolution (4K/8K). To reduce the computational and memory cost, current methods first split the large input into local patches and then merge the SR patches into the output. These methods adaptively allocate a subnet for each patch. Quantization is a very important technique for network acceleration and has been used to design the subnets. Current methods train an MLP bit selector to determine the propoer bit for each layer. However, they uniformly sample subnets for training, making simple subnets overfitted and complicated subnets underfitted. Therefore, the trained bit selector fails to determine the optimal bit. Apart from this, the introduced bit selector brings additional cost to each layer of the SR network. In this paper, we propose a novel method named Content-Aware Bit Mapping (CABM), which can remove the bit selector without any performance loss. CABM also learns a bit selector for each layer during training. After training, we analyze the relation between the edge information of an input patch and the bit of each layer. We observe that the edge information can be an effective metric for the selected bit. Therefore, we design a strategy to build an Edge-to-Bit lookup table that maps the edge score of a patch to the bit of each layer during inference. The bit configuration of SR network can be determined by the lookup tables of all layers. Our strategy can find better bit configuration, resulting in more efficient mixed precision networks. We conduct detailed experiments to demonstrate the generalization ability of our method. The code will be released.
****
Decoupling MaxLogit for Out-of-Distribution Detection
Zihan Zhang, Xiang Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3388-3397
In machine learning, it is often observed that standard training outputs anomalously high confidence for both in-distribution (ID) and out-of-distribution (OOD) data. Thus, the ability to detect OOD samples is critical to the model deployment. An essential step for OOD detection is post-hoc scoring. MaxLogit is one of the simplest scoring functions which uses the maximum logits as OOD score. To provide a new viewpoint to study the logit-based scoring function, we reformulate the logit into cosine similarity and logit norm and propose to use MaxCosine and MaxNorm. We empirically find that MaxCosine is a core factor in the effectiveness of MaxLogit. And the performance of MaxLogit is encumbered by MaxNorm. To tackle the problem, we propose the Decoupling MaxLogit (DML) for flexibility to balance MaxCosine and MaxNorm. To further embody the core of our method, we extend DML to DML+ based on the new insights that fewer hard samples and compact feature space are the key components to make logit-based methods effective. We demonstrate the effectiveness of our logit-based OOD detection methods on CIFAR-10, CIFAR-100 and ImageNet and establish state-of-the-art performance.
****
ProphNet: Efficient Agent-Centric Motion Forecasting With Anchor-Informed Proposals
Xishun Wang, Tong Su, Fang Da, Xiaodong Yang; Proceedings of the IEEE/CVF Confer

ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21995-22003
Motion forecasting is a key module in an autonomous driving system. Due to the heterogeneous nature of multi-sourced input, multimodality in agent behavior, and low latency required by onboard deployment, this task is notoriously challenging. To cope with these difficulties, this paper proposes a novel agent-centric model with anchor-informed proposals for efficient multimodal motion forecasting. We design a modality-agnostic strategy to concisely encode the complex input in a unified manner. We generate diverse proposals, fused with anchors bearing goal-oriented context, to induce multimodal prediction that covers a wide range of future trajectories. The network architecture is highly uniform and succinct, leading to an efficient model amenable for real-world deployment. Experiments reveal that our agent-centric network compares favorably with the state-of-the-art methods in prediction accuracy, while achieving scene-centric level inference latency.
****
Generalizing Dataset Distillation via Deep Generative Prior
George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3739-3748
Dataset Distillation aims to distill an entire dataset's knowledge into a few synthetic images. The idea is to synthesize a small number of synthetic data points that, when given to a learning algorithm as training data, result in a model approximating one trained on the original data. Despite a recent upsurge of progress in the field, existing dataset distillation methods fail to generalize to new architectures and scale to high-resolution datasets. To overcome the above issues, we propose to use the learned prior from pre-trained deep generative models to synthesize the distilled data. To achieve this, we present a new optimization algorithm that distills a large number of images into a few intermediate feature vectors in the generative model's latent space. Our method augments existing techniques, significantly improving cross-architecture generalization in all settings.
****
Few-Shot Class-Incremental Learning via Class-Aware Bilateral Distillation
Linglan Zhao, Jing Lu, Yunlu Xu, Zhanzhan Cheng, Dashan Guo, Yi Niu, Xiangzhong Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11838-11847
Few-Shot Class-Incremental Learning (FSCIL) aims to continually learn novel classes based on only few training samples, which poses a more challenging task than the well-studied Class-Incremental Learning (CIL) due to data scarcity. While knowledge distillation, a prevailing technique in CIL, can alleviate the catastrophic forgetting of older classes by regularizing outputs between current and previous model, it fails to consider the overfitting risk of novel classes in FSCIL. To adapt the powerful distillation technique for FSCIL, we propose a novel distillation structure, by taking the unique challenge of overfitting into account. Concretely, we draw knowledge from two complementary teachers. One is the model trained on abundant data from base classes that carries rich general knowledge, which can be leveraged for easing the overfitting of current novel classes. The other is the updated model from last incremental session that contains the adapted knowledge of previous novel classes, which is used for alleviating their forgetting. To combine the guidances, an adaptive strategy conditioned on the class-wise semantic similarities is introduced. Besides, for better preserving base class knowledge when accommodating novel concepts, we adopt a two-branch network with an attention-based aggregation module to dynamically merge predictions from two complementary branches. Extensive experiments on 3 popular FSCIL datasets: mini-ImageNet, CIFAR100 and CUB200 validate the effectiveness of our method by surpassing existing works by a significant margin.
****
Adaptive Patch Deformation for Textureless-Resilient Multi-View Stereo
Yuesong Wang, Zhaojie Zeng, Tao Guan, Wei Yang, Zhuo Chen, Wenkai Liu, Luoyuan Xu, Yawei Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt

ern Recognition (CVPR), 2023, pp. 1621-1630
In recent years, deep learning-based approaches have shown great strength in mul
ti-view stereo because of their outstanding ability to extract robust visual fea
tures. However, most learning-based methods need to build the cost volume and in
crease the receptive field enormously to get a satisfactory result when dealing
with large-scale textureless regions, consequently leading to prohibitive memory
 consumption. To ensure both memory-friendly and textureless-resilient, we innov
atively transplant the spirit of deformable convolution from deep learning into
the traditional PatchMatch-based method. Specifically, for each pixel with match
ing ambiguity (termed unreliable pixel), we adaptively deform the patch centered
 on it to extend the receptive field until covering enough correlative reliable
pixels (without matching ambiguity) that serve as anchors. When performing Patch
Match, constrained by the anchor pixels, the matching cost of an unreliable pixe
l is guaranteed to reach the global minimum at the correct depth and therefore i
ncreases the robustness of multi-view stereo significantly. To detect more ancho
r pixels to ensure better adaptive patch deformation, we propose to evaluate the
 matching ambiguity of a certain pixel by checking the convergence of the estima
ted depth as optimization proceeds. As a result, our method achieves state-of-th
e-art performance on ETH3D and Tanks and Temples while preserving low memory con
sumption.
****
Detection of Out-of-Distribution Samples Using Binary Neuron Activation Patterns
Bart■omiej Olber, Krystian Radlak, Adam Popowicz, Michal Szczepankiewicz, Krysti
an Chachu■a; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2023, pp. 3378-3387
Deep neural networks (DNN) have outstanding performance in various applications.
 Despite numerous efforts of the research community, out-of-distribution (OOD) s
amples remain a significant limitation of DNN classifiers. The ability to identi
fy previously unseen inputs as novel is crucial in safety-critical applications
such as self-driving cars, unmanned aerial vehicles, and robots. Existing approa
ches to detect OOD samples treat a DNN as a black box and evaluate the confidenc
e score of the output predictions. Unfortunately, this method frequently fails,
because DNNs are not trained to reduce their confidence for OOD inputs. In this
work, we introduce a novel method for OOD detection. Our method is motivated by
theoretical analysis of neuron activation patterns (NAP) in ReLU-based architect
ures. The proposed method does not introduce a high computational overhead due t
o the binary representation of the activation patterns extracted from convolutio
nal layers. The extensive empirical evaluation proves its high performance on va
rious DNN architectures and seven image datasets.
****
SeaThru-NeRF: Neural Radiance Fields in Scattering Media
Deborah Levy, Amit Peleg, Naama Pearl, Dan Rosenbaum, Derya Akkaynak, Simon Korm
an, Tali Treibitz; Proceedings of the IEEE/CVF Conference on Computer Vision and
 Pattern Recognition (CVPR), 2023, pp. 56-65
Research on neural radiance fields (NeRFs) for novel view generation is explodin
g with new models and extensions. However, a question that remains unanswered is
 what happens in underwater or foggy scenes where the medium strongly influences
 the appearance of objects. Thus far, NeRF and its variants have ignored these c
ases. However, since the NeRF framework is based on volumetric rendering, it has
 inherent capability to account for the medium's effects, once modeled appropria
tely. We develop a new rendering model for NeRFs in scattering media, which is b
ased on the SeaThru image formation model, and suggest a suitable architecture f
or learning both scene information and medium parameters. We demonstrate the str
ength of our method using simulated and real-world scenes, correctly rendering n
ovel photorealistic views underwater. Even more excitingly, we can render clear
views of these scenes, removing the medium between the camera and the scene and
reconstructing the appearance and depth of far objects, which are severely occlu
ded by the medium. Our code and unique datasets are available on the project's w
ebsite.
****

Learning Multi-Modal Class-Specific Tokens for Weakly Supervised Dense Object Lo calization
Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, Dan Xu; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 3, pp. 19596-19605
Weakly supervised dense object localization (WSDOL) relies generally on Class Ac tivation Mapping (CAM), which exploits the correlation between the class weights of the image classifier and the pixel-level features. Due to the limited abilit y to address intra-class variations, the image classifier cannot properly associ ate the pixel features, leading to inaccurate dense localization maps. In this p aper, we propose to explicitly construct multi-modal class representations by le veraging the Contrastive Language-Image Pre-training (CLIP), to guide dense loca lization. More specifically, we propose a unified transformer framework to learn two-modalities of class-specific tokens, i.e., class-specific visual and textua l tokens. The former captures semantics from the target visual data while the la tter exploits the class-related language priors from CLIP, providing complementa ry information to better perceive the intra-class diversities. In addition, we p ropose to enrich the multi-modal class-specific tokens with sample-specific cont exts comprising visual context and image-language context. This enables more ada ptive class representation learning, which further facilitates dense localizatio n. Extensive experiments show the superiority of the proposed method for WSDOL o n two multi-label datasets, i.e., PASCAL VOC and MS COCO, and one single-label d ataset, i.e., OpenImages. Our dense localization maps also lead to the state-of- the-art weakly supervised semantic segmentation (WSSS) results on PASCAL VOC and MS COCO.
****
Learning To Dub Movies via Hierarchical Prosody Models
Gaoxiang Cong, Liang Li, Yuankai Qi, Zheng-Jun Zha, Qi Wu, Wenyu Wang, Bin Jiang , Ming-Hsuan Yang, Qingming Huang; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2023, pp. 14687-14697
Given a piece of text, a video clip and a reference audio, the movie dubbing (al so known as visual voice clone, V2C) task aims to generate speeches that match t he speaker's emotion presented in the video using the desired speaker voice as r eference. V2C is more challenging than conventional text-to-speech tasks as it a dditionally requires the generated speech to exactly match the varying emotions and speaking speed presented in the video. Unlike previous works, we propose a n ovel movie dubbing architecture to tackle these problems via hierarchical prosod y modeling, which bridges the visual information to corresponding speech prosody from three aspects: lip, face, and scene. Specifically, we align lip movement t o the speech duration, and convey facial expression to speech energy and pitch v ia attention mechanism based on valence and arousal representations inspired by the psychology findings. Moreover, we design an emotion booster to capture the a tmosphere from global video scenes. All these embeddings are used together to ge nerate mel-spectrogram, which is then converted into speech waves by an existing vocoder. Extensive experimental results on the V2C and Chem benchmark datasets demonstrate the favourable performance of the proposed method. The code and trai ned models will be made available at https://github.com/GalaxyCong/HPMDubbing.
****
DiffusionRig: Learning Personalized Priors for Facial Appearance Editing
Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, Xiuming Zhang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2023, pp. 12736-12746
We address the problem of learning person-specific facial priors from a small nu mber (e.g., 20) of portrait photos of the same person. This enables us to edit t his specific person's facial appearance, such as expression and lighting, while preserving their identity and high-frequency facial details. Key to our approach , which we dub DiffusionRig, is a diffusion model conditioned on, or "rigged by, " crude 3D face models estimated from single in-the-wild images by an off-the-sh elf estimator. On a high level, DiffusionRig learns to map simplistic renderings of 3D face models to realistic photos of a given person. Specifically, Diffusio

nRig is trained in two stages: It first learns generic facial priors from a larg
e-scale face dataset and then person-specific priors from a small portrait photo
 collection of the person of interest. By learning the CGI-to-photo mapping with
 such personalized priors, DiffusionRig can "rig" the lighting, facial expressio
n, head pose, etc. of a portrait photo, conditioned only on coarse 3D models whi
le preserving this person's identity and other high-frequency characteristics. Q
ualitative and quantitative experiments show that DiffusionRig outperforms exist
ing approaches in both identity preservation and photorealism. Please see the pr
oject website: https://diffusionrig.github.io for the supplemental material, vid
eo, code, and data.
****
Delving StyleGAN Inversion for Image Editing: A Foundation Latent Space Viewpoin
t
Hongyu Liu, Yibing Song, Qifeng Chen; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10072-10082
GAN inversion and editing via StyleGAN maps an input image into the embedding sp
aces (W, W^+, and F) to simultaneously maintain image fidelity and meaningful ma
nipulation. From latent space W to extended latent space W^+ to feature space F
in StyleGAN, the editability of GAN inversion decreases while its reconstruction
 quality increases. Recent GAN inversion methods typically explore W^+ and F rat
her than W to improve reconstruction fidelity while maintaining editability. As
W^+ and F are derived from W that is essentially the foundation latent space of
StyleGAN, these GAN inversion methods focusing on W^+ and F spaces could be impr
oved by stepping back to W. In this work, we propose to first obtain the proper
latent code in foundation latent space W. We introduce contrastive learning to a
lign W and the image space for proper latent code discovery. Then, we leverage a
 cross-attention encoder to transform the obtained latent code in W into W^+ and
 F, accordingly. Our experiments show that our exploration of the foundation lat
ent space W improves the representation ability of latent codes in W^+ and featu
res in F, which yields state-of-the-art reconstruction fidelity and editability
results on the standard benchmarks. Project page: https://kumapowerliu.github.io
/CLCAE.
****
MixMAE: Mixed and Masked Autoencoder for Efficient Pretraining of Hierarchical V
ision Transformers
Jihao Liu, Xin Huang, Jinliang Zheng, Yu Liu, Hongsheng Li; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
6252-6261
In this paper, we propose Mixed and Masked AutoEncoder (MixMAE), a simple but ef
ficient pretraining method that is applicable to various hierarchical Vision Tra
nsformers. Existing masked image modeling (MIM) methods for hierarchical Vision
Transformers replace a random subset of input tokens with a special [MASK] symbo
l and aim at reconstructing original image tokens from the corrupted image. Howe
ver, we find that using the [MASK] symbol greatly slows down the training and ca
uses pretraining-finetuning inconsistency, due to the large masking ratio (e.g.,
 60% in SimMIM). On the other hand, MAE does not introduce [MASK] tokens at its
encoder at all but is not applicable for hierarchical Vision Transformers. To so
lve the issue and accelerate the pretraining of hierarchical models, we replace
the masked tokens of one image with visible tokens of another image, i.e., creat
ing a mixed image. We then conduct dual reconstruction to reconstruct the two or
iginal images from the mixed input, which significantly improves efficiency. Whi
le MixMAE can be applied to various hierarchical Transformers, this paper explor
es using Swin Transformer with a large window size and scales up to huge model s
ize (to reach 600M parameters). Empirical results demonstrate that MixMAE can le
arn high-quality visual representations efficiently. Notably, MixMAE with Swin-B
/W14 achieves 85.1% top-1 accuracy on ImageNet-1K by pretraining for 600 epochs.
 Besides, its transfer performances on the other 6 datasets show that MixMAE has
 better FLOPs / performance tradeoff than previous popular MIM methods.
****
Human Pose Estimation in Extremely Low-Light Conditions

Sohyun Lee, Jaesung Rim, Boseung Jeong, Geonu Kim, Byungju Woo, Haechan Lee, Sun ghyun Cho, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 704-714
We study human pose estimation in extremely low-light images. This task is chall enging due to the difficulty of collecting real low-light images with accurate l abels, and severely corrupted inputs that degrade prediction quality significant ly. To address the first issue, we develop a dedicated camera system and build a new dataset of real low-light images with accurate pose labels. Thanks to our c amera system, each low-light image in our dataset is coupled with an aligned wel l-lit image, which enables accurate pose labeling and is used as privileged info rmation during training. We also propose a new model and a new training strategy that fully exploit the privileged information to learn representation insensiti ve to lighting conditions. Our method demonstrates outstanding performance on re al extremely low-light images, and extensive analyses validate that both of our model and dataset contribute to the success.
****
EventNeRF: Neural Radiance Fields From a Single Colour Event Camera
Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, Vladislav Golyanik; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4992-5002
Asynchronously operating event cameras find many applications due to their high dynamic range, vanishingly low motion blur, low latency and low data bandwidth. The field saw remarkable progress during the last few years, and existing event-based 3D reconstruction approaches recover sparse point clouds of the scene. How ever, such sparsity is a limiting factor in many cases, especially in computer v ision and graphics, that has not been addressed satisfactorily so far. According ly, this paper proposes the first approach for 3D-consistent, dense and photorea listic novel view synthesis using just a single colour event stream as input. At its core is a neural radiance field trained entirely in a self-supervised manne r from events while preserving the original resolution of the colour event chann els. Next, our ray sampling strategy is tailored to events and allows for data-e fficient training. At test, our method produces results in the RGB space at unpr ecedented quality. We evaluate our method qualitatively and numerically on sever al challenging synthetic and real scenes and show that it produces significantly denser and more visually appealing renderings than the existing methods. We als o demonstrate robustness in challenging scenarios with fast motion and under low lighting conditions. We release the newly recorded dataset and our source code to facilitate the research field, see https://4dqv.mpi-inf.mpg.de/EventNeRF.
****
Neighborhood Attention Transformer
Ali Hassani, Steven Walton, Jiachen Li, Shen Li, Humphrey Shi; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 6185-6194
We present Neighborhood Attention (NA), the first efficient and scalable sliding window attention mechanism for vision. NA is a pixel-wise operation, localizing self attention (SA) to the nearest neighboring pixels, and therefore enjoys a l inear time and space complexity compared to the quadratic complexity of SA. The sliding window pattern allows NA's receptive field to grow without needing extra pixel shifts, and preserves translational equivariance, unlike Swin Transformer 's Window Self Attention (WSA). We develop NATTEN (Neighborhood Attention Extens ion), a Python package with efficient C++ and CUDA kernels, which allows NA to r un up to 40% faster than Swin's WSA while using up to 25% less memory. We furthe r present Neighborhood Attention Transformer (NAT), a new hierarchical transform er design based on NA that boosts image classification and downstream vision per formance. Experimental results on NAT are competitive; NAT-Tiny reaches 83.2% to p-1 accuracy on ImageNet, 51.4% mAP on MS-COCO and 48.4% mIoU on ADE20K, which i s 1.9% ImageNet accuracy, 1.0% COCO mAP, and 2.6% ADE20K mIoU improvement over a Swin model with similar size. To support more research based on sliding window attention, we open source our project and release our checkpoints.
****

Enlarging Instance-Specific and Class-Specific Information for Open-Set Action Recognition

Jun Cen, Shiwei Zhang, Xiang Wang, Yixuan Pei, Zhiwu Qing, Yingya Zhang, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15295-15304

Open-set action recognition is to reject unknown human action cases which are out of the distribution of the training set. Existing methods mainly focus on learning better uncertainty scores but dismiss the importance of feature representations. We find that features with richer semantic diversity can significantly improve the open-set performance under the same uncertainty scores. In this paper, we begin with analyzing the feature representation behavior in the open-set action recognition (OSAR) problem based on the information bottleneck (IB) theory, and propose to enlarge the instance-specific (IS) and class-specific (CS) information contained in the feature for better performance. To this end, a novel Prototypical Similarity Learning (PSL) framework is proposed to keep the instance variance within the same class to retain more IS information. Besides, we notice that unknown samples sharing similar appearances to known samples are easily misclassified as known classes. To alleviate this issue, video shuffling is further introduced in our PSL to learn distinct temporal information between original and shuffled samples, which we find enlarges the CS information. Extensive experiments demonstrate that the proposed PSL can significantly boost both the open-set and closed-set performance and achieves state-of-the-art results on multiple benchmarks. Code is available at https://github.com/Jun-CEN/PSL.

****

Decoupled Semantic Prototypes Enable Learning From Diverse Annotation Types for Semi-Weakly Segmentation in Expert-Driven Domains

Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15495-15506

A vast amount of images and pixel-wise annotations allowed our community to build scalable segmentation solutions for natural domains. However, the transfer to expert-driven domains like microscopy applications or medical healthcare remains difficult as domain experts are a critical factor due to their limited availability for providing pixel-wise annotations. To enable affordable segmentation solutions for such domains, we need training strategies which can simultaneously handle diverse annotation types and are not bound to costly pixel-wise annotations. In this work, we analyze existing training algorithms towards their flexibility for different annotation types and scalability to small annotation regimes. We conduct an extensive evaluation in the challenging domain of organelle segmentation and find that existing semi- and semi-weakly supervised training algorithms are not able to fully exploit diverse annotation types. Driven by our findings, we introduce Decoupled Semantic Prototypes (DSP) as a training method for semantic segmentation which enables learning from annotation types as diverse as image-level-, point-, bounding box-, and pixel-wise annotations and which leads to remarkable accuracy gains over existing solutions for semi-weakly segmentation.

****

Progressive Spatio-Temporal Alignment for Efficient Event-Based Motion Estimation

Xueyan Huang, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1537-1546

In this paper, we propose an efficient event-based motion estimation framework for various motion models. Different from previous works, we design a progressive event-to-map alignment scheme and utilize the spatio-temporal correlations to align events. In detail, we progressively align sampled events in an event batch to the time-surface map and obtain the updated motion model by minimizing a novel time-surface loss. In addition, a dynamic batch size strategy is applied to adaptively adjust the batch size so that all events in the batch are consistent with the current motion model. Our framework has three advantages: a) the progressive scheme refines motion parameters iteratively, achieving accurate motion estimation; b) within one iteration, only a small portion of events are involved in

optimization, which greatly reduces the total runtime; c) the dynamic batch size strategy ensures that the constant velocity assumption always holds. We conduct comprehensive experiments to evaluate our framework on challenging high-speed scenes with three motion models: rotational, homography, and 6-DOF models. Experimental results demonstrate that our framework achieves state-of-the-art estimation accuracy and efficiency.
****
Trap Attention: Monocular Depth Estimation With Manual Traps
Chao Ning, Hongping Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5033-5043
Predicting a high quality depth map from a single image is a challenging task, because it exists infinite possibility to project a 2D scene to the corresponding 3D scene. Recently, some studies introduced multi-head attention (MHA) modules to perform long-range interaction, which have shown significant progress in regressing the depth maps.The main functions of MHA can be loosely summarized to capture long-distance information and report the attention map by the relationship between pixels. However, due to the quadratic complexity of MHA, these methods can not leverage MHA to compute depth features in high resolution with an appropriate computational complexity. In this paper, we exploit a depth-wise convolution to obtain long-range information, and propose a novel trap attention, which sets some traps on the extended space for each pixel, and forms the attention mechanism by the feature retention ratio of convolution window, resulting in that the quadratic computational complexity can be converted to linear form. Then we build an encoder-decoder trap depth estimation network, which introduces a vision transformer as the encoder, and uses the trap attention to estimate the depth from single image in the decoder. Extensive experimental results demonstrate that our proposed network can outperform the state-of-the-art methods in monocular depth estimation on datasets NYU Depth-v2 and KITTI, with significantly reduced number of parameters. Code is available at: https://github.com/ICSResearch/TrapAttention.
****
Iterative Next Boundary Detection for Instance Segmentation of Tree Rings in Microscopy Images of Shrub Cross Sections
Alexander Gillert, Giulia Resente, Alba Anadon-Rosell, Martin Wilmking, Uwe Freiherr von Lukas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14540-14548
We address the problem of detecting tree rings in microscopy images of shrub cross sections. This can be regarded as a special case of the instance segmentation task with several unique challenges such as the concentric circular ring shape of the objects and high precision requirements that result in inadequate performance of existing methods. We propose a new iterative method which we term Iterative Next Boundary Detection (INBD). It intuitively models the natural growth direction, starting from the center of the shrub cross section and detecting the next ring boundary in each iteration step. In our experiments, INBD shows superior performance to generic instance segmentation methods and is the only one with a built-in notion of chronological order. Our dataset and source code are available at http://github.com/alexander-g/INBD.
****
Learning and Aggregating Lane Graphs for Urban Automated Driving
Martin Büchner, Jannik Zürn, Ion-George Todoran, Abhinav Valada, Wolfram Burgard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13415-13424
Lane graph estimation is an essential and highly challenging task in automated driving and HD map learning. Existing methods using either onboard or aerial imagery struggle with complex lane topologies, out-of-distribution scenarios, or significant occlusions in the image space. Moreover, merging overlapping lane graphs to obtain consistent largescale graphs remains difficult. To overcome these challenges, we propose a novel bottom-up approach to lane graph estimation from aerial imagery that aggregates multiple overlapping graphs into a single consistent graph. Due to its modular design, our method allows us to address two compleme

ntary tasks: predicting ego-respective successor lane graphs from arbitrary vehi
cle positions using a graph neural network and aggregating these predictions int
o a consistent global lane graph. Extensive experiments on a large-scale lane gr
aph dataset demonstrate that our approach yields highly accurate lane graphs, ev
en in regions with severe occlusions. The presented approach to graph aggregatio
n proves to eliminate inconsistent predictions while increasing the overall grap
h quality. We make our large-scale urban lane graph dataset and code publicly av
ailable at http://urbanlanegraph.cs.uni-freiburg.de.
****
Universal Instance Perception As Object Discovery and Retrieval
Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, Huchuan Lu; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 15325-15336
All instance perception tasks aim at finding certain objects specified by some q
ueries such as category names, language expressions, and target annotations, but
 this complete field has been split into multiple independent subtasks. In this
work, we present a universal instance perception model of the next generation, t
ermed UNINEXT. UNINEXT reformulates diverse instance perception tasks into a uni
fied object discovery and retrieval paradigm and can flexibly perceive different
 types of objects by simply changing the input prompts. This unified formulation
 brings the following benefits: (1) enormous data from different tasks and label
 vocabularies can be exploited for jointly training general instance-level repre
sentations, which is especially beneficial for tasks lacking in training data. (
2) the unified model is parameter-efficient and can save redundant computation w
hen handling multiple tasks simultaneously. UNINEXT shows superior performance o
n 20 challenging benchmarks from 10 instance-level tasks including classical ima
ge-level tasks (object detection and instance segmentation), vision-and-language
 tasks (referring expression comprehension and segmentation), and six video-leve
l object tracking tasks. Code is available at https://github.com/MasterBin-IIAU/
UNINEXT.
****
GlassesGAN: Eyewear Personalization Using Synthetic Appearance Discovery and Tar
geted Subspace Modeling
Richard Plesh, Peter Peer, Vitomir Struc; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16847-16857
We present GlassesGAN, a novel image editing framework for custom design of glas
ses, that sets a new standard in terms of output-image quality, edit realism, an
d continuous multi-style edit capability. To facilitate the editing process with
 GlassesGAN, we propose a Targeted Subspace Modelling (TSM) procedure that, base
d on a novel mechanism for (synthetic) appearance discovery in the latent space
of a pre-trained GAN generator, constructs an eyeglasses-specific (latent) subsp
ace that the editing framework can utilize. Additionally, we also introduce an a
ppearance-constrained subspace initialization (SI) technique that centers the la
tent representation of the given input image in the well-defined part of the con
structed subspace to improve the reliability of the learned edits. We test Glass
esGAN on two (diverse) high-resolution datasets (CelebA-HQ and SiblingsDB-HQf) a
nd compare it to three state-of-the-art baselines, i.e., InterfaceGAN, GANSpace,
 and MaskGAN. The reported results show that GlassesGAN convincingly outperforms
 all competing techniques, while offering functionality (e.g., fine-grained mult
i-style editing) not available with any of the competitors. The source code for
GlassesGAN is made publicly available.
****
Representing Volumetric Videos As Dynamic MLP Maps
Sida Peng, Yunzhi Yan, Qing Shuai, Hujun Bao, Xiaowei Zhou; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
4252-4262
This paper introduces a novel representation of volumetric videos for real-time
view synthesis of dynamic scenes. Recent advances in neural scene representation
s demonstrate their remarkable capability to model and render complex static sce
nes, but extending them to represent dynamic scenes is not straightforward due t

o their slow rendering speed or high storage cost. To solve this problem, our ke
y idea is to represent the radiance field of each frame as a set of shallow MLP
networks whose parameters are stored in 2D grids, called MLP maps, and dynamical
ly predicted by a 2D CNN decoder shared by all frames. Representing 3D scenes wi
th shallow MLPs significantly improves the rendering speed, while dynamically pr
edicting MLP parameters with a shared 2D CNN instead of explicitly storing them
leads to low storage cost. Experiments show that the proposed approach achieves
state-of-the-art rendering quality on the NHR and ZJU-MoCap datasets, while bein
g efficient for real-time rendering with a speed of 41.7 fps for 512 x 512 image
s on an RTX 3090 GPU. The code is available at https://zju3dv.github.io/mlp_maps
/.
****
Deep Hashing With Minimal-Distance-Separated Hash Centers
Liangdao Wang, Yan Pan, Cong Liu, Hanjiang Lai, Jian Yin, Ye Liu; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 23455-23464
Deep hashing is an appealing approach for large-scale image retrieval. Most exis
ting supervised deep hashing methods learn hash functions using pairwise or trip
le image similarities in randomly sampled mini-batches. They suffer from low tra
ining efficiency, insufficient coverage of data distribution, and pair imbalance
 problems. Recently, central similarity quantization (CSQ) attacks the above pro
blems by using "hash centers" as a global similarity metric, which encourages th
e hash codes of similar images to approach their common hash center and distance
 themselves from other hash centers. Although achieving SOTA retrieval performan
ce, CSQ falls short of a worst-case guarantee on the minimal distance between it
s constructed hash centers, i.e. the hash centers can be arbitrarily close. This
 paper presents an optimization method that finds hash centers with a constraint
 on the minimal distance between any pair of hash centers, which is non-trivial
due to the non-convex nature of the problem. More importantly, we adopt the Gilb
ert-Varshamov bound from coding theory, which helps us to obtain a large minimal
 distance while ensuring the empirical feasibility of our optimization approach.
 With these clearly-separated hash centers, each is assigned to one image class,
 we propose several effective loss functions to train deep hashing networks. Ext
ensive experiments on three datasets for image retrieval demonstrate that the pr
oposed method achieves superior retrieval performance over the state-of-the-art
deep hashing methods.
****
Video-Text As Game Players: Hierarchical Banzhaf Interaction for Cross-Modal Rep
resentation Learning
Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, L
i Yuan, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2023, pp. 2472-2482
Contrastive learning-based video-language representation learning approaches, e.
g., CLIP, have achieved outstanding performance, which pursue semantic interacti
on upon pre-defined video-text pairs. To clarify this coarse-grained global inte
raction and move a step further, we have to encounter challenging shell-breaking
 interactions for fine-grained cross-modal learning. In this paper, we creativel
y model video-text as game players with multivariate cooperative game theory to
wisely handle the uncertainty during fine-grained semantic interaction with dive
rse granularity, flexible combination, and vague intensity. Concretely, we propo
se Hierarchical Banzhaf Interaction (HBI) to value possible correspondence betwe
en video frames and text words for sensitive and explainable cross-modal contras
t. To efficiently realize the cooperative game of multiple video frames and mult
iple text words, the proposed method clusters the original video frames (text wo
rds) and computes the Banzhaf Interaction between the merged tokens. By stacking
 token merge modules, we achieve cooperative games at different semantic levels.
 Extensive experiments on commonly used text-video retrieval and video-question
answering benchmarks with superior performances justify the efficacy of our HBI.
 More encouragingly, it can also serve as a visualization tool to promote the un
derstanding of cross-modal interaction, which may have a far-reaching impact on

the community. Project page is available at https://jpthu17.github.io/HBI/.
****
VL-SAT: Visual-Linguistic Semantics Assisted Training for 3D Semantic Scene Grap
h Prediction in Point Cloud
Ziqin Wang, Bowen Cheng, Lichen Zhao, Dong Xu, Yang Tang, Lu Sheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21560-21569
The task of 3D semantic scene graph (3DSSG) prediction in the point cloud is challenging since (1) the 3D point cloud only captures geometric structures with limited semantics compared to 2D images, and (2) long-tailed relation distribution inherently hinders the learning of unbiased prediction. Since 2D images provide rich semantics and scene graphs are in nature coped with languages, in this study, we propose Visual-Linguistic Semantics Assisted Training (VL-SAT) scheme that can significantly empower 3DSSG prediction models with discrimination about long-tailed and ambiguous semantic relations. The key idea is to train a powerful multi-modal oracle model to assist the 3D model. This oracle learns reliable structural representations based on semantics from vision, language, and 3D geometry, and its benefits can be heterogeneously passed to the 3D model during the training stage. By effectively utilizing visual-linguistic semantics in training, our VL-SAT can significantly boost common 3DSSG prediction models, such as SGFN and SGGpoint, only with 3D inputs in the inference stage, especially when dealing with tail relation triplets. Comprehensive evaluations and ablation studies on the 3DSSG dataset have validated the effectiveness of the proposed scheme. Code is available at https://github.com/wz7in/CVPR2023-VLSAT.
****
Learning Emotion Representations From Verbal and Nonverbal Communication
Sitao Zhang, Yimu Pan, James Z. Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18993-19004
Emotion understanding is an essential but highly challenging component of artificial general intelligence. The absence of extensive annotated datasets has significantly impeded advancements in this field. We present EmotionCLIP, the first pre-training paradigm to extract visual emotion representations from verbal and nonverbal communication using only uncurated data. Compared to numerical labels or descriptions used in previous methods, communication naturally contains emotion information. Furthermore, acquiring emotion representations from communication is more congruent with the human learning process. We guide EmotionCLIP to attend to nonverbal emotion cues through subject-aware context encoding and verbal emotion cues using sentiment-guided contrastive learning. Extensive experiments validate the effectiveness and transferability of EmotionCLIP. Using merely linear-probe evaluation protocol, EmotionCLIP outperforms the state-of-the-art supervised visual emotion recognition methods and rivals many multimodal approaches across various benchmarks. We anticipate that the advent of EmotionCLIP will address the prevailing issue of data scarcity in emotion understanding, thereby fostering progress in related domains. The code and pre-trained models are available at https://github.com/Xeaver/EmotionCLIP.
****
Transferable Adversarial Attacks on Vision Transformers With Token Gradient Regu
larization
Jianping Zhang, Yizhan Huang, Weibin Wu, Michael R. Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16415-16424
Vision transformers (ViTs) have been successfully deployed in a variety of computer vision tasks, but they are still vulnerable to adversarial samples. Transfer-based attacks use a local model to generate adversarial samples and directly transfer them to attack a target black-box model. The high efficiency of transfer-based attacks makes it a severe security threat to ViT-based applications. Therefore, it is vital to design effective transfer-based attacks to identify the deficiencies of ViTs beforehand in security-sensitive scenarios. Existing efforts generally focus on regularizing the input gradients to stabilize the updated direction of adversarial samples. However, the variance of the back-propagated gradi

ents in intermediate blocks of ViTs may still be large, which may make the gener ated adversarial samples focus on some model-specific features and get stuck in poor local optima. To overcome the shortcomings of existing approaches, we propo se the Token Gradient Regularization (TGR) method. According to the structural c haracteristics of ViTs, TGR reduces the variance of the back-propagated gradient in each internal block of ViTs in a token-wise manner and utilizes the regulari zed gradient to generate adversarial samples. Extensive experiments on attacking both ViTs and CNNs confirm the superiority of our approach. Notably, compared t o the state-of-the-art transfer-based attacks, our TGR offers a performance impr ovement of 8.8 % on average.
****
MCF: Mutual Correction Framework for Semi-Supervised Medical Image Segmentation
Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, Xinbo Gao; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1 5651-15660
Semi-supervised learning is a promising method for medical image segmentation un der limited annotation. However, the model cognitive bias impairs the segmentati on performance, especially for edge regions. Furthermore, current mainstream sem i-supervised medical image segmentation (SSMIS) methods lack designs to handle m odel bias. The neural network has a strong learning ability, but the cognitive b ias will gradually deepen during the training, and it is difficult to correct it self. We propose a novel mutual correction framework (MCF) to explore network bi as correction and improve the performance of SSMIS. Inspired by the plain contra st idea, MCF introduces two different subnets to explore and utilize the discrep ancies between subnets to correct cognitive bias of the model. More concretely, a contrastive difference review (CDR) module is proposed to find out inconsisten t prediction regions and perform a review training. Additionally, a dynamic comp etitive pseudo-label generation (DCPLG) module is proposed to evaluate the perfo rmance of subnets in real-time, dynamically selecting more reliable pseudo-label s. Experimental results on two medical image databases with different modalities (CT and MRI) show that our method achieves superior performance compared to sev eral state-of-the-art methods. The code will be available at https://github.com/ WYC-321/MCF.
****
Blur Interpolation Transformer for Real-World Motion From Blur
Zhihang Zhong, Mingdeng Cao, Xiang Ji, Yinqiang Zheng, Imari Sato; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 3, pp. 5713-5723
This paper studies the challenging problem of recovering motion from blur, also known as joint deblurring and interpolation or blur temporal super-resolution. T he challenges are twofold: 1) the current methods still leave considerable room for improvement in terms of visual quality even on the synthetic dataset, and 2) poor generalization to real-world data. To this end, we propose a blur interpol ation transformer (BiT) to effectively unravel the underlying temporal correlati on encoded in blur. Based on multi-scale residual Swin transformer blocks, we in troduce dual-end temporal supervision and temporally symmetric ensembling strate gies to generate effective features for time-varying motion rendering. In additi on, we design a hybrid camera system to collect the first real-world dataset of one-to-many blur-sharp video pairs. Experimental results show that BiT has a sig nificant gain over the state-of-the-art methods on the public dataset Adobe240. Besides, the proposed real-world dataset effectively helps the model generalize well to real blurry scenarios. Code and data are available at https://github.com /zzh-tech/BiT.
****
Rethinking Few-Shot Medical Segmentation: A Vector Quantization View
Shiqi Huang, Tingfa Xu, Ning Shen, Feng Mu, Jianan Li; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3072- 3081
The existing few-shot medical segmentation networks share the same practice that the more prototypes, the better performance. This phenomenon can be theoretical

ly interpreted in Vector Quantization (VQ) view: the more prototypes, the more c
lusters are separated from pixel-wise feature points distributed over the full s
pace. However, as we further think about few-shot segmentation with this perspec
tive, it is found that the clusterization of feature points and the adaptation t
o unseen tasks have not received enough attention. Motivated by the observation,
 we propose a learning VQ mechanism consisting of grid-format VQ (GFVQ), self-or
ganized VQ (SOVQ) and residual oriented VQ (ROVQ). To be specific, GFVQ generate
s the prototype matrix by averaging square grids over the spatial extent, which
uniformly quantizes the local details; SOVQ adaptively assigns the feature point
s to different local classes and creates a new representation space where the le
arnable local prototypes are updated with a global view; ROVQ introduces residua
l information to fine-tune the aforementioned learned local prototypes without r
e-training, which benefits the generalization performance for the irrelevance to
 the training task. We empirically show that our VQ framework yields the state-o
f-the-art performance over abdomen, cardiac and prostate MRI datasets and expect
 this work will provoke a rethink of the current few-shot medical segmentation m
odel design. Our code will soon be publicly available.
****
Event-Based Shape From Polarization
Manasi Muglikar, Leonard Bauersfeld, Diederik Paul Moeys, Davide Scaramuzza; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 1547-1556
State-of-the-art solutions for Shape-from-Polarization (SfP) suffer from a speed
-resolution tradeoff: they either sacrifice the number of polarization angles me
asured or necessitate lengthy acquisition times due to framerate constraints, th
us compromising either accuracy or latency. We tackle this tradeoff using event
cameras. Event cameras operate at microseconds resolution with negligible motion
 blur, and output a continuous stream of events that precisely measures how ligh
t changes over time asynchronously. We propose a setup that consists of a linear
 polarizer rotating at high speeds in front of an event camera. Our method uses
the continuous event stream caused by the rotation to reconstruct relative inten
sities at multiple polarizer angles. Experiments demonstrate that our method out
performs physics-based baselines using frames, reducing the MAE by 25% in synthe
tic and real-world datasets. In the real world, we observe, however, that the ch
allenging conditions (i.e., when few events are generated) harm the performance
of physics-based solutions. To overcome this, we propose a learning-based approa
ch that learns to estimate surface normals even at low event-rates, improving th
e physics-based approach by 52% on the real world dataset. The proposed system a
chieves an acquisition speed equivalent to 50 fps (>twice the framerate of the c
ommercial polarization sensor) while retaining the spatial resolution of 1MP. Ou
r evaluation is based on the first large-scale dataset for event-based SfP.
****
Architectural Backdoors in Neural Networks
Mikel Bober-Irizar, Ilia Shumailov, Yiren Zhao, Robert Mullins, Nicolas Papernot
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2023, pp. 24595-24604
Machine learning is vulnerable to adversarial manipulation. Previous literature
has demonstrated that at the training stage attackers can manipulate data (Gu et
 al.) and data sampling procedures (Shumailov et al.) to control model behaviour
. A common attack goal is to plant backdoors i.e. force the victim model to lear
n to recognise a trigger known only by the adversary. In this paper, we introduc
e a new class of backdoor attacks that hide inside model architectures i.e. in t
he inductive bias of the functions used to train. These backdoors are simple to
implement, for instance by publishing open-source code for a backdoored model ar
chitecture that others will reuse unknowingly. We demonstrate that model archite
ctural backdoors represent a real threat and, unlike other approaches, can survi
ve a complete re-training from scratch. We formalise the main construction princ
iples behind architectural backdoors, such as a connection between the input and
 the output, and describe some possible protections against them. We evaluate ou
r attacks on computer vision benchmarks of different scales and demonstrate the

underlying vulnerability is pervasive in a variety of common training settings.
****
ARO-Net: Learning Implicit Fields From Anchored Radial Observations
Yizhi Wang, Zeyu Huang, Ariel Shamir, Hui Huang, Hao Zhang, Ruizhen Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3572-3581
We introduce anchored radial observations (ARO), a novel shape encoding for learning implicit field representation of 3D shapes that is category-agnostic and generalizable amid significant shape variations. The main idea behind our work is to reason about shapes through partial observations from a set of viewpoints, called anchors. We develop a general and unified shape representation by employing a fixed set of anchors, via Fibonacci sampling, and designing a coordinate-based deep neural network to predict the occupancy value of a query point in space. Differently from prior neural implicit models that use global shape feature, our shape encoder operates on contextual, query-specific features. To predict point occupancy, locally observed shape information from the perspective of the anchors surrounding the input query point are encoded and aggregated through an attention module, before implicit decoding is performed. We demonstrate the quality and generality of our network, coined ARO-Net, on surface reconstruction from sparse point clouds, with tests on novel and unseen object categories, "one-shape" training, and comparisons to state-of-the-art neural and classical methods for reconstruction and tessellation.
****
All in One: Exploring Unified Video-Language Pre-Training
Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6598-6608
Mainstream Video-Language Pre-training models consist of three parts, a video encoder, a text encoder, and a video-text fusion Transformer. They pursue better performance via utilizing heavier unimodal encoders or multimodal fusion Transformers, resulting in increased parameters with lower efficiency in downstream tasks. In this work, we for the first time introduce an end-to-end video-language model, namely all-in-one Transformer, that embeds raw video and textual signals into joint representations using a unified backbone architecture. We argue that the unique temporal information of video data turns out to be a key barrier hindering the design of a modality-agnostic Transformer. To overcome the challenge, we introduce a novel and effective token rolling operation to encode temporal representations from video clips in a non-parametric manner. The careful design enables the representation learning of both video-text multimodal inputs and unimodal inputs using a unified backbone model. Our pre-trained all-in-one Transformer is transferred to various downstream video-text tasks after fine-tuning, including text-video retrieval, video-question answering, multiple choice and visual commonsense reasoning. State-of-the-art performances with the minimal model FLOPs on nine datasets demonstrate the superiority of our method compared to the competitive counterparts.
****
Parametric Implicit Face Representation for Audio-Driven Facial Reenactment
Ricong Huang, Peiwen Lai, Yipeng Qin, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12759-12768
Audio-driven facial reenactment is a crucial technique that has a range of applications in film-making, virtual avatars and video conferences. Existing works either employ explicit intermediate face representations (e.g., 2D facial landmarks or 3D face models) or implicit ones (e.g., Neural Radiance Fields), thus suffering from the trade-offs between interpretability and expressive power, hence between controllability and quality of the results. In this work, we break these trade-offs with our novel parametric implicit face representation and propose a novel audio-driven facial reenactment framework that is both controllable and can generate high-quality talking heads. Specifically, our parametric implicit repr

esentation parameterizes the implicit representation with interpretable paramete rs of 3D face models, thereby taking the best of both explicit and implicit meth ods. In addition, we propose several new techniques to improve the three compone nts of our framework, including i) incorporating contextual information into the audio-to-expression parameters encoding; ii) using conditional image synthesis to parameterize the implicit representation and implementing it with an innovati ve tri-plane structure for efficient learning; iii) formulating facial reenactme nt as a conditional image inpainting problem and proposing a novel data augmenta tion technique to improve model generalizability. Extensive experiments demonstr ate that our method can generate more realistic results than previous methods wi th greater fidelity to the identities and talking styles of speakers.
****
Semantic Human Parsing via Scalable Semantic Transfer Over Multiple Label Domain s
Jie Yang, Chaoqun Wang, Zhen Li, Junle Wang, Ruimao Zhang; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1 9424-19433
This paper presents Scalable Semantic Transfer (SST), a novel training paradigm, to explore how to leverage the mutual benefits of the data from different label domains (i.e. various levels of label granularity) to train a powerful human pa rsing network. In practice, two common application scenarios are addressed, term ed universal parsing and dedicated parsing, where the former aims to learn homog eneous human representations from multiple label domains and switch predictions by only using different segmentation heads, and the latter aims to learn a speci fic domain prediction while distilling the semantic knowledge from other domains . The proposed SST has the following appealing benefits: (1) it can capably serv e as an effective training scheme to embed semantic associations of human body p arts from multiple label domains into the human representation learning process; (2) it is an extensible semantic transfer framework without predetermining the overall relations of multiple label domains, which allows continuously adding hu man parsing datasets to promote the training. (3) the relevant modules are only used for auxiliary training and can be removed during inference, eliminating the extra reasoning cost. Experimental results demonstrate SST can effectively achi eve promising universal human parsing performance as well as impressive improvem ents compared to its counterparts on three human parsing benchmarks (i.e., PASCA L-Person-Part, ATR, and CIHP). Code is available at https://github.com/yangjie-c v/SST.
****
Making Vision Transformers Efficient From a Token Sparsification View
Shuning Chang, Pichao Wang, Ming Lin, Fan Wang, David Junhao Zhang, Rong Jin, Mi ke Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2023, pp. 6195-6205
The quadratic computational complexity to the number of tokens limits the practi cal applications of Vision Transformers (ViTs). Several works propose to prune r edundant tokens to achieve efficient ViTs. However, these methods generally suff er from (i) dramatic accuracy drops, (ii) application difficulty in the local vi sion transformer, and (iii) non-general-purpose networks for downstream tasks. I n this work, we propose a novel Semantic Token ViT (STViT), for efficient global and local vision transformers, which can also be revised to serve as backbone f or downstream tasks. The semantic tokens represent cluster centers, and they are initialized by pooling image tokens in space and recovered by attention, which can adaptively represent global or local semantic information. Due to the cluste r properties, a few semantic tokens can attain the same effect as vast image tok ens, for both global and local vision transformers. For instance, only 16 semant ic tokens on DeiT-(Tiny,Small,Base) can achieve the same accuracy with more than 100% inference speed improvement and nearly 60% FLOPs reduction; on Swin-(Tiny, Small,Base), we can employ 16 semantic tokens in each window to further speed it up by around 20% with slight accuracy increase. Besides great success in image classification, we also extend our method to video recognition. In addition, we design a STViT-R(ecovery) network to restore the detailed spatial information ba

sed on the STViT, making it work for downstream tasks, which is powerless for previous token sparsification methods. Experiments demonstrate that our method can achieve competitive results compared to the original networks in object detection and instance segmentation, with over 30% FLOPs reduction for backbone.
****
GEN: Pushing the Limits of Softmax-Based Out-of-Distribution Detection
Xixi Liu, Yaroslava Lochman, Christopher Zach; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23946-23955
Out-of-distribution (OOD) detection has been extensively studied in order to successfully deploy neural networks, in particular, for safety-critical applications. Moreover, performing OOD detection on large-scale datasets is closer to reality, but is also more challenging. Several approaches need to either access the training data for score design or expose models to outliers during training. Some post-hoc methods are able to avoid the aforementioned constraints, but are less competitive. In this work, we propose Generalized ENtropy score (GEN), a simple but effective entropy-based score function, which can be applied to any pre-trained softmax-based classifier. Its performance is demonstrated on the large-scale ImageNet-1k OOD detection benchmark. It consistently improves the average AUROC across six commonly-used CNN-based and visual transformer classifiers over a number of state-of-the-art post-hoc methods. The average AUROC improvement is at least 3.5%. Furthermore, we used GEN on top of feature-based enhancing methods as well as methods using training statistics to further improve the OOD detection performance. The code is available at: https://github.com/XixiLiu95/GEN.
****
RefCLIP: A Universal Teacher for Weakly Supervised Referring Expression Comprehension
Lei Jin, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Annan Shu, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2681-2690
Referring Expression Comprehension (REC) is a task of grounding the referent based on an expression, and its development is greatly limited by expensive instance-level annotations. Most existing weakly supervised methods are built based on two-stage detection networks, which are computationally expensive. In this paper, we resort to the efficient one-stage detector and propose a novel weakly supervised model called RefCLIP. Specifically, RefCLIP redefines weakly supervised REC as an anchor-text matching problem, which can avoid the complex post-processing in existing methods. To achieve weakly supervised learning, we introduce anchor-based contrastive loss to optimize RefCLIP via numerous anchor-text pairs. Based on RefCLIP, we further propose the first model-agnostic weakly supervised training scheme for existing REC models, where RefCLIP acts as a mature teacher to generate pseudo-labels for teaching common REC models. With our careful designs, this scheme can even help existing REC models achieve better weakly supervised performance than RefCLIP, e.g., TransVG and SimREC. To validate our approaches, we conduct extensive experiments on four REC benchmarks, i.e., RefCOCO, RefCOCO+, RefCOCOg and ReferItGame. Experimental results not only report our significant performance gains over existing weakly supervised models, e.g., +24.87% on RefCOCO, but also show the 5x faster inference speed. Project: https://refclip.github.io.
****
VILA: Learning Image Aesthetics From User Comments With Vision-Language Pretraining
Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, Feng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10041-10051
Assessing the aesthetics of an image is challenging, as it is influenced by multiple factors including composition, color, style, and high-level semantics. Existing image aesthetic assessment (IAA) methods primarily rely on human-labeled rating scores, which oversimplify the visual aesthetic information that humans perceive. Conversely, user comments offer more comprehensive information and are a more natural way to express human opinions and preferences regarding image aesth

etics. In light of this, we propose learning image aesthetics from user comments, and exploring vision-language pretraining methods to learn multimodal aesthetic representations. Specifically, we pretrain an image-text encoder-decoder model with image-comment pairs, using contrastive and generative objectives to learn rich and generic aesthetic semantics without human labels. To efficiently adapt the pretrained model for downstream IAA tasks, we further propose a lightweight rank-based adapter that employs text as an anchor to learn the aesthetic ranking concept. Our results show that our pretrained aesthetic vision-language model outperforms prior works on image aesthetic captioning over the AVA-Captions dataset, and it has powerful zero-shot capability for aesthetic tasks such as zero-shot style classification and zero-shot IAA, surpassing many supervised baselines. With only minimal finetuning parameters using the proposed adapter module, our model achieves state-of-the-art IAA performance over the AVA dataset.
****
Learnable Skeleton-Aware 3D Point Cloud Sampling
Cheng Wen, Baosheng Yu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17671-17681
Point cloud sampling is crucial for efficient large-scale point cloud analysis, where learning-to-sample methods have recently received increasing attention from the community for jointly training with downstream tasks. However, the above-mentioned task-specific sampling methods usually fail to explore the geometries of objects in an explicit manner. In this paper, we introduce a new skeleton-aware learning-to-sample method by learning object skeletons as the prior knowledge to preserve the object geometry and topology information during sampling. Specifically, without labor-intensive annotations per object category, we first learn category-agnostic object skeletons via the medial axis transform definition in an unsupervised manner. With object skeleton, we then evaluate the histogram of the local feature size as the prior knowledge to formulate skeleton-aware sampling from a probabilistic perspective. Additionally, the proposed skeleton-aware sampling pipeline with the task network is thus end-to-end trainable by exploring the reparameterization trick. Extensive experiments on three popular downstream tasks, point cloud classification, retrieval, and reconstruction, demonstrate the effectiveness of the proposed method for efficient point cloud analysis.
****
Boundary-Enhanced Co-Training for Weakly Supervised Semantic Segmentation
Shenghai Rong, Bohai Tu, Zilei Wang, Junjie Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19574-19584
The existing weakly supervised semantic segmentation (WSSS) methods pay much attention to generating accurate and complete class activation maps (CAMs) as pseudo-labels, while ignoring the importance of training the segmentation networks. In this work, we observe that there is an inconsistency between the quality of the pseudo-labels in CAMs and the performance of the final segmentation model, and the mislabeled pixels mainly lie on the boundary areas. Inspired by these findings, we argue that the focus of WSSS should be shifted to robust learning given the noisy pseudo-labels, and further propose a boundary-enhanced co-training (BECO) method for training the segmentation model. To be specific, we first propose to use a co-training paradigm with two interactive networks to improve the learning of uncertain pixels. Then we propose a boundary-enhanced strategy to boost the prediction of difficult boundary areas, which utilizes reliable predictions to construct artificial boundaries. Benefiting from the design of co-training and boundary enhancement, our method can achieve promising segmentation performance for different CAMs. Extensive experiments on PASCAL VOC 2012 and MS COCO 2014 validate the superiority of our BECO over other state-of-the-art methods.
****
Re-IQA: Unsupervised Learning for Image Quality Assessment in the Wild
Avinab Saha, Sandeep Mishra, Alan C. Bovik; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5846-5855
Automatic Perceptual Image Quality Assessment is a challenging problem that impacts billions of internet, and social media users daily. To advance research in this field, we propose a Mixture of Experts approach to train two separate encode

rs to learn high-level content and low-level image quality features in an unsupe rvised setting. The unique novelty of our approach is its ability to generate low-level representations of image quality that are complementary to high-level features representing image content. We refer to the framework used to train the two encoders as Re-IQA. For Image Quality Assessment in the Wild, we deploy the complementary low and high-level image representations obtained from the Re-IQA framework to train a linear regression model, which is used to map the image representations to the ground truth quality scores, refer Figure 1. Our method achieves state-of-the-art performance on multiple large-scale image quality assessment databases containing both real and synthetic distortions, demonstrating how deep neural networks can be trained in an unsupervised setting to produce perceptually relevant representations. We conclude from our experiments that the low and high-level features obtained are indeed complementary and positively impact the performance of the linear regressor. A public release of all the codes associated with this work will be made available on GitHub.
****
Procedure-Aware Pretraining for Instructional Video Understanding
Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, Juan Carlos Niebles; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10727-10738
Our goal is to learn a video representation that is useful for downstream procedure understanding tasks in instructional videos. Due to the small amount of available annotations, a key challenge in procedure understanding is to be able to extract from unlabeled videos the procedural knowledge such as the identity of the task (e.g., 'make latte'), its steps (e.g., 'pour milk'), or the potential next steps given partial progress in its execution. Our main insight is that instructional videos depict sequences of steps that repeat between instances of the same or different tasks, and that this structure can be well represented by a Procedural Knowledge Graph (PKG), where nodes are discrete steps and edges connect steps that occur sequentially in the instructional activities. This graph can then be used to generate pseudo labels to train a video representation that encodes the procedural knowledge in a more accessible form to generalize to multiple procedure understanding tasks. We build a PKG by combining information from a text-based procedural knowledge database and an unlabeled instructional video corpus and then use it to generate training pseudo labels with four novel pre-training objectives. We call this PKG-based pre-training procedure and the resulting model Paprika, Procedure-Aware PRe-training for Instructional Knowledge Acquisition. We evaluate Paprika on COIN and CrossTask for procedure understanding tasks such as task recognition, step recognition, and step forecasting. Paprika yields a video representation that improves over the state of the art: up to 11.23% gains in accuracy in 12 evaluation settings. Implementation is available at https://github.com/salesforce/paprika.
****
Sample-Level Multi-View Graph Clustering
Yuze Tan, Yixi Liu, Shudong Huang, Wentao Feng, Jiancheng Lv; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23966-23975
Multi-view clustering have hitherto been studied due to their effectiveness in dealing with heterogeneous data. Despite the empirical success made by recent works, there still exists several severe challenges. Particularly, previous multi-view clustering algorithms seldom consider the topological structure in data, which is essential for clustering data on manifold. Moreover, existing methods cannot fully consistency the consistency of local structures between different views as they explore the clustering structure in a view-wise manner. In this paper, we propose to exploit the implied data manifold by learning the topological structure of data. Besides, considering that the consistency of multiple views is manifested in the generally similar local structure while the inconsistent structures are minority, we further explore the intersections of multiple views in the sample level such that the cross-view consistency can be better maintained. We model the above concerns in a unified framework and design an efficient algorithm

to solve the corresponding optimization problem. Experimental results on various multi-view datasets certificate the effectiveness of the proposed method and verify its superiority over other SOTA approaches.
****

Fine-Grained Audible Video Description
Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, Yu Qiao, Yiran Zhong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10585-10596
We explore a new task for audio-visual-language modeling called fine-grained audible video description (FAVD). It aims to provide detailed textual descriptions for the given audible videos, including the appearance and spatial locations of each object, the actions of moving objects, and the sounds in videos. Existing visual-language modeling tasks often concentrate on visual cues in videos while undervaluing the language and audio modalities. On the other hand, FAVD requires not only audio-visual-language modeling skills but also paragraph-level language generation abilities. We construct the first fine-grained audible video description benchmark (FAVDBench) to facilitate this research. For each video clip, we first provide a one-sentence summary of the video, ie, the caption, followed by 4-6 sentences describing the visual details and 1-2 audio-related descriptions at the end. The descriptions are provided in both English and Chinese. We create two new metrics for this task: an EntityScore to gauge the completeness of entities in the visual descriptions, and an AudioScore to assess the audio descriptions. As a preliminary approach to this task, we propose an audio-visual-language transformer that extends existing video captioning model with an additional audio branch. We combine the masked language modeling and auto-regressive language modeling losses to optimize our model so that it can produce paragraph-level descriptions. We illustrate the efficiency of our model in audio-visual-language modeling by evaluating it against the proposed benchmark using both conventional captioning metrics and our proposed metrics. We further put our benchmark to the test in video generation models, demonstrating that employing fine-grained video descriptions can create more intricate videos than using captions. Code and dataset are available at https://github.com/OpenNLPLab/FAVDBench. Our online benchmark is available at www.avlbench.opennlplab.cn.
****

3D Semantic Segmentation in the Wild: Learning Generalized Models for Adverse-Condition Point Clouds
Aoran Xiao, Jiaxing Huang, Weihao Xuan, Ruijie Ren, Kangcheng Liu, Dayan Guan, Abdulmotaleb El Saddik, Shijian Lu, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9382-9392
Robust point cloud parsing under all-weather conditions is crucial to level-5 autonomy in autonomous driving. However, how to learn a universal 3D semantic segmentation (3DSS) model is largely neglected as most existing benchmarks are dominated by point clouds captured under normal weather. We introduce SemanticSTF, an adverse-weather point cloud dataset that provides dense point-level annotations and allows to study 3DSS under various adverse weather conditions. We investigate universal 3DSS modeling with two tasks: 1) domain adaptive 3DSS that adapts from normal-weather data to adverse-weather data; 2) domain generalized 3DSS that learns a generalizable model from normal-weather data. Our studies reveal the challenge while existing 3DSS methods encounter adverse-weather data, showing the great value of SemanticSTF in steering the future endeavor along this very meaningful research direction. In addition, we design a domain randomization technique that alternatively randomizes the geometry styles of point clouds and aggregates their encoded embeddings, ultimately leading to a generalizable model that effectively improves 3DSS under various adverse weather. The SemanticSTF and related codes are available at https://github.com/xiaoaoran/SemanticSTF.
****

Catch Missing Details: Image Reconstruction With Frequency Augmented Variational Autoencoder
Xinmiao Lin, Yikang Li, Jenhao Hsiao, Chiuman Ho, Yu Kong; Proceedings of the IE

EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1
736-1745
The popular VQ-VAE models reconstruct images through learning a discrete codeboo
k but suffer from a significant issue in the rapid quality degradation of image
reconstruction as the compression rate rises. One major reason is that a higher
compression rate induces more loss of visual signals on the higher frequency spe
ctrum, which reflect the details on pixel space. In this paper, a Frequency Comp
lement Module (FCM) architecture is proposed to capture the missing frequency in
formation for enhancing reconstruction quality. The FCM can be easily incorporat
ed into the VQ-VAE structure, and we refer to the new model as Frequancy Augment
ed VAE (FA-VAE). In addition, a Dynamic Spectrum Loss (DSL) is introduced to gui
de the FCMs to balance between various frequencies dynamically for optimal recon
struction. FA-VAE is further extended to the text-to-image synthesis task, and a
 Cross-attention Autoregressive Transformer (CAT) is proposed to obtain more pre
cise semantic attributes in texts. Extensive reconstruction experiments with dif
ferent compression rates are conducted on several benchmark datasets, and the re
sults demonstrate that the proposed FA-VAE is able to restore more faithfully th
e details compared to SOTA methods. CAT also shows improved generation quality w
ith better image-text semantic alignment.
****
RaBit: Parametric Modeling of 3D Biped Cartoon Characters With a Topological-Con
sistent Dataset
Zhongjin Luo, Shengcai Cai, Jinguo Dong, Ruibo Ming, Liangdong Qiu, Xiaohang Zha
n, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2023, pp. 12825-12835
Assisting people in efficiently producing visually plausible 3D characters has a
lways been a fundamental research topic in computer vision and computer graphics
. Recent learning-based approaches have achieved unprecedented accuracy and effi
ciency in the area of 3D real human digitization. However, none of the prior wor
ks focus on modeling 3D biped cartoon characters, which are also in great demand
 in gaming and filming. In this paper, we introduce 3DBiCar, the first large-sca
le dataset of 3D biped cartoon characters, and RaBit, the corresponding parametr
ic model. Our dataset contains 1,500 topologically consistent high-quality 3D te
xtured models which are manually crafted by professional artists. Built upon the
 data, RaBit is thus designed with a SMPL-like linear blend shape model and a St
yleGAN-based neural UV-texture generator, simultaneously expressing the shape, p
ose, and texture. To demonstrate the practicality of 3DBiCar and RaBit, various
applications are conducted, including single-view reconstruction, sketch-based m
odeling, and 3D cartoon animation. For the single-view reconstruction setting, w
e find a straightforward global mapping from input images to the output UV-based
 texture maps tends to lose detailed appearances of some local parts (e.g., nose
, ears). Thus, a part-sensitive texture reasoner is adopted to make all importan
t local areas perceived. Experiments further demonstrate the effectiveness of ou
r method both qualitatively and quantitatively. 3DBiCar and RaBit are available
at gaplab.cuhk.edu.cn/projects/RaBit.
****
Next3D: Generative Neural Texture Rasterization for 3D-Aware Head Avatars
Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, Yeb
in Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2023, pp. 20991-21002
3D-aware generative adversarial networks (GANs) synthesize high-fidelity and mul
ti-view-consistent facial images using only collections of single-view 2D imager
y. Towards fine-grained control over facial attributes, recent efforts incorpora
te 3D Morphable Face Model (3DMM) to describe deformation in generative radiance
 fields either explicitly or implicitly. Explicit methods provide fine-grained e
xpression control but cannot handle topological changes caused by hair and acces
sories, while implicit ones can model varied topologies but have limited general
ization caused by the unconstrained deformation fields. We propose a novel 3D GA
N framework for unsupervised learning of generative, high-quality and 3D-consist
ent facial avatars from unstructured 2D images. To achieve both deformation accu

racy and topological flexibility, we propose a 3D representation called Generati
ve Texture-Rasterized Tri-planes. The proposed representation learns Generative
Neural Textures on top of parametric mesh templates and then projects them into
three orthogonal-viewed feature planes through rasterization, forming a tri-plan
e feature representation for volume rendering. In this way, we combine both fine
-grained expression control of mesh-guided explicit deformation and the flexibil
ity of implicit volumetric representation. We further propose specific modules f
or modeling mouth interior which is not taken into account by 3DMM. Our method d
emonstrates state-of-the-art 3Daware synthesis quality and animation ability thr
ough extensive experiments. Furthermore, serving as 3D prior, our animatable 3D
representation boosts multiple applications including one-shot facial avatars an
d 3D-aware stylization.
****
Uni3D: A Unified Baseline for Multi-Dataset 3D Object Detection
Bo Zhang, Jiakang Yuan, Botian Shi, Tao Chen, Yikang Li, Yu Qiao; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 9253-9262
Current 3D object detection models follow a single dataset-specific training and
 testing paradigm, which often faces a serious detection accuracy drop when they
 are directly deployed in another dataset. In this paper, we study the task of t
raining a unified 3D detector from multiple datasets. We observe that this appea
rs to be a challenging task, which is mainly due to that these datasets present
substantial data-level differences and taxonomy-level variations caused by diffe
rent LiDAR types and data acquisition standards. Inspired by such observation, w
e present a Uni3D which leverages a simple data-level correction operation and a
 designed semantic-level coupling-and-recoupling module to alleviate the unavoid
able data-level and taxonomy-level differences, respectively. Our method is simp
le and easily combined with many 3D object detection baselines such as PV-RCNN a
nd Voxel-RCNN, enabling them to effectively learn from multiple off-the-shelf 3D
 datasets to obtain more discriminative and generalizable representations. Exper
iments are conducted on many dataset consolidation settings. Their results demon
strate that Uni3D exceeds a series of individual detectors trained on a single d
ataset, with a 1.04x parameter increase over a selected baseline detector. We ex
pect this work will inspire the research of 3D generalization since it will push
 the limits of perceptual performance. Our code is available at: https://github.
com/PJLab-ADG/3DTrans
****
Linking Garment With Person via Semantically Associated Landmarks for Virtual Tr
y-On
Keyu Yan, Tingwei Gao, Hui Zhang, Chengjun Xie; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17194-17204
In this paper, a novel virtual try-on algorithm, dubbed SAL-VTON, is proposed, w
hich links the garment with the person via semantically associated landmarks to
alleviate misalignment. The semantically associated landmarks are a series of la
ndmark pairs with the same local semantics on the in-shop garment image and the
try-on image. Based on the semantically associated landmarks, SAL-VTON effective
ly models the local semantic association between garment and person, making up f
or the misalignment in the overall deformation of the garment. The outcome is ac
hieved with a three-stage framework: 1) the semantically associated landmarks ar
e estimated using the landmark localization model; 2) taking the landmarks as in
put, the warping model explicitly associates the corresponding parts of the garm
ent and person for obtaining the local flow, thus refining the alignment in the
global flow; 3) finally, a generator consumes the landmarks to better capture lo
cal semantics and control the try-on results.Moreover, we propose a new landmark
 dataset with a unified labelling rule of landmarks for diverse styles of garmen
ts. Extensive experimental results on popular datasets demonstrate that SAL-VTON
 can handle misalignment and outperform state-of-the-art methods both qualitativ
ely and quantitatively. The dataset is available on https://modelscope.cn/datase
ts/damo/SAL-HG/summary.
****

ACR: Attention Collaboration-Based Regressor for Arbitrary Two-Hand Reconstruction
Zhengdi Yu, Shaoli Huang, Chen Fang, Toby P. Breckon, Jue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12955-12964
Reconstructing two hands from monocular RGB images is challenging due to frequent occlusion and mutual confusion. Existing methods mainly learn an entangled representation to encode two interacting hands, which are incredibly fragile to impaired interaction, such as truncated hands, separate hands, or external occlusion. This paper presents ACR (Attention Collaboration-based Regressor), which makes the first attempt to reconstruct hands in arbitrary scenarios. To achieve this, ACR explicitly mitigates interdependencies between hands and between parts by leveraging center and part-based attention for feature extraction. However, reducing interdependence helps release the input constraint while weakening the mutual reasoning about reconstructing the interacting hands. Thus, based on center attention, ACR also learns cross-hand prior that handle the interacting hands better. We evaluate our method on various types of hand reconstruction datasets. Our method significantly outperforms the best interacting-hand approaches on the InterHand2.6M dataset while yielding comparable performance with the state-of-the-art single-hand methods on the FreiHand dataset. More qualitative results on in-the-wild and hand-object interaction datasets and web images/videos further demonstrate the effectiveness of our approach for arbitrary hand reconstruction. Our code is available at https://github.com/ZhengdiYu/Arbitrary-Hands-3D-Reconstruction
****
Rotation-Invariant Transformer for Point Cloud Matching
Hao Yu, Zheng Qin, Ji Hou, Mahdi Saleh, Dongsheng Li, Benjamin Busam, Slobodan Ilic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5384-5393
The intrinsic rotation invariance lies at the core of matching point clouds with handcrafted descriptors. However, it is widely despised by recent deep matchers that obtain the rotation invariance extrinsically via data augmentation. As the finite number of augmented rotations can never span the continuous SO(3) space, these methods usually show instability when facing rotations that are rarely seen. To this end, we introduce RoITr, a Rotation-Invariant Transformer to cope with the pose variations in the point cloud matching task. We contribute both on the local and global levels. Starting from the local level, we introduce an attention mechanism embedded with Point Pair Feature (PPF)-based coordinates to describe the pose-invariant geometry, upon which a novel attention-based encoder-decoder architecture is constructed. We further propose a global transformer with rotation-invariant cross-frame spatial awareness learned by the self-attention mechanism, which significantly improves the feature distinctiveness and makes the model robust with respect to the low overlap. Experiments are conducted on both the rigid and non-rigid public benchmarks, where RoITr outperforms all the state-of-the-art models by a considerable margin in the low-overlapping scenarios. Especially when the rotations are enlarged on the challenging 3DLoMatch benchmark, RoITr surpasses the existing methods by at least 13 and 5 percentage points in terms of Inlier Ratio and Registration Recall, respectively.
****
Devil's on the Edges: Selective Quad Attention for Scene Graph Generation
Deunsol Jung, Sanghyun Kim, Won Hwa Kim, Minsu Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18664-18674
Scene graph generation aims to construct a semantic graph structure from an image such that its nodes and edges respectively represent objects and their relationships. One of the major challenges for the task lies in the presence of distracting objects and relationships in images; contextual reasoning is strongly distracted by irrelevant objects or backgrounds and, more importantly, a vast number of irrelevant candidate relations. To tackle the issue, we propose the Selective Quad Attention Network (SQUAT) that learns to select relevant object pairs and

disambiguate them via diverse contextual interactions. SQUAT consists of two main components: edge selection and quad attention. The edge selection module selects relevant object pairs, i.e., edges in the scene graph, which helps contextual reasoning, and the quad attention module then updates the edge features using both edge-to-node and edge-to-edge cross-attentions to capture contextual information between objects and object pairs. Experiments demonstrate the strong performance and robustness of SQUAT, achieving the state of the art on the Visual Genome and Open Images v6 benchmarks.
****
NIFF: Alleviating Forgetting in Generalized Few-Shot Object Detection via Neural Instance Feature Forging
Karim Guirguis, Johannes Meier, George Eskandar, Matthias Kayser, Bin Yang, Jürgen Beyerer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24193-24202
Privacy and memory are two recurring themes in a broad conversation about the societal impact of AI. These concerns arise from the need for huge amounts of data to train deep neural networks. A promise of Generalized Few-shot Object Detection (G-FSOD), a learning paradigm in AI, is to alleviate the need for collecting abundant training samples of novel classes we wish to detect by leveraging prior knowledge from old classes (i.e., base classes). G-FSOD strives to learn these novel classes while alleviating catastrophic forgetting of the base classes. However, existing approaches assume that the base images are accessible, an assumption that does not hold when sharing and storing data is problematic. In this work, we propose the first data-free knowledge distillation (DFKD) approach for G-FSOD that leverages the statistics of the region of interest (RoI) features from the base model to forge instance-level features without accessing the base images. Our contribution is three-fold: (1) we design a standalone lightweight generator with (2) class-wise heads (3) to generate and replay diverse instance-level base features to the RoI head while finetuning on the novel data. This stands in contrast to standard DFKD approaches in image classification, which invert the entire network to generate base images. Moreover, we make careful design choices in the novel finetuning pipeline to regularize the model. We show that our approach can dramatically reduce the base memory requirements, all while setting a new standard for G-FSOD on the challenging MS-COCO and PASCAL-VOC benchmarks.
****
Habitat-Matterport 3D Semantics Dataset
Karmesh Yadav, Ram Ramrakhya, Santhosh Kumar Ramakrishnan, Theo Gervet, John Turner, Aaron Gokaslan, Noah Maestre, Angel Xuan Chang, Dhruv Batra, Manolis Savva, Alexander William Clegg, Devendra Singh Chaplot; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4927-4936
We present the Habitat-Matterport 3D Semantics (HM3DSEM) dataset. HM3DSEM is the largest dataset of 3D real-world spaces with densely annotated semantics that is currently available to the academic community. It consists of 142,646 object instance annotations across 216 3D spaces and 3,100 rooms within those spaces. The scale, quality, and diversity of object annotations far exceed those of prior datasets. A key difference setting apart HM3DSEM from other datasets is the use of texture information to annotate pixel-accurate object boundaries. We demonstrate the effectiveness of HM3DSEM dataset for the Object Goal Navigation task using different methods. Policies trained using HM3DSEM perform outperform those trained on prior datasets. Introduction of HM3DSEM in the Habitat ObjectNav Challenge lead to an increase in participation from 400 submissions in 2021 to 1022 submissions in 2022. Project page: https://aihabitat.org/datasets/hm3d-semantics/
****
Post-Processing Temporal Action Detection
Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18837-18845
Existing Temporal Action Detection (TAD) methods typically take a pre-processing step in converting an input varying-length video into a fixed-length snippet representation sequence, before temporal boundary estimation and action classifica

tion. This pre-processing step would temporally downsample the video, reducing the inference resolution and hampering the detection performance in the original temporal resolution. In essence, this is due to a temporal quantization error introduced during the resolution downsampling and recovery. This could negatively impact the TAD performance, but is largely ignored by existing methods. To address this problem, in this work we introduce a novel model-agnostic post-processing method without model redesign and retraining. Specifically, we model the start and end points of action instances with a Gaussian distribution for enabling temporal boundary inference at a sub-snippet level. We further introduce an efficient Taylor-expansion based approximation, dubbed as Gaussian Approximated Post-processing (GAP). Extensive experiments demonstrate that our GAP can consistently improve a wide variety of pre-trained off-the-shelf TAD models on the challenging ActivityNet (+0.2% 0.7% in average mAP) and THUMOS (+0.2% 0.5% in average mAP) benchmarks. Such performance gains are already significant and highly comparable to those achieved by novel model designs. Also, GAP can be integrated with model training for further performance gain. Importantly, GAP enables lower temporal resolutions for more efficient inference, facilitating low-resource applications. The code is available in https://github.com/sauradip/GAP.
****
ConZIC: Controllable Zero-Shot Image Captioning by Sampling-Based Polishing
Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, Zhengjue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23465-23476
Zero-shot capability has been considered as a new revolution of deep learning, letting machines work on tasks without curated training data. As a good start and the only existing outcome of zero-shot image captioning (IC), ZeroCap abandons supervised training and sequentially searching every word in the caption using the knowledge of large-scale pre-trained models. Though effective, its autoregressive generation and gradient-directed searching mechanism limit the diversity of captions and inference speed, respectively. Moreover, ZeroCap does not consider the controllability issue of zero-shot IC. To move forward, we propose a framework for Controllable Zero-shot IC, named ConZIC. The core of ConZIC is a novel sampling-based non-autoregressive language model named GibbsBERT, which can generate and continuously polish every word. Extensive quantitative and qualitative results demonstrate the superior performance of our proposed ConZIC for both zero-shot IC and controllable zero-shot IC. Especially, ConZIC achieves about 5x faster generation speed than ZeroCap, and about 1.5x higher diversity scores, with accurate generation given different control signals.
****
EDGE: Editable Dance Generation From Music
Jonathan Tseng, Rodrigo Castellon, Karen Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 448-458
Dance is an important human art form, but creating new dances can be difficult and time-consuming. In this work, we introduce Editable Dance GEneration (EDGE), a state-of-the-art method for editable dance generation that is capable of creating realistic, physically-plausible dances while remaining faithful to the input music. EDGE uses a transformer-based diffusion model paired with Jukebox, a strong music feature extractor, and confers powerful editing capabilities well-suited to dance, including joint-wise conditioning, and in-betweening. We introduce a new metric for physical plausibility, and evaluate dance quality generated by our method extensively through (1) multiple quantitative metrics on physical plausibility, alignment, and diversity benchmarks, and more importantly, (2) a large-scale user study, demonstrating a significant improvement over previous state-of-the-art methods. Qualitative samples from our model can be found at our website.
****
Curricular Contrastive Regularization for Physics-Aware Single Image Dehazing
Yu Zheng, Jiahui Zhan, Shengfeng He, Junyu Dong, Yong Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5785-5794

Considering the ill-posed nature, contrastive regularization has been developed for single image dehazing, introducing the information from negative images as a lower bound. However, the contrastive samples are nonconsensual, as the negatives are usually represented distantly from the clear (i.e., positive) image, leaving the solution space still under-constricted. Moreover, the interpretability of deep dehazing models is underexplored towards the physics of the hazing process. In this paper, we propose a novel curricular contrastive regularization targeted at a consensual contrastive space as opposed to a non-consensual one. Our negatives, which provide better lower-bound constraints, can be assembled from 1) the hazy image, and 2) corresponding restorations by other existing methods. Further, due to the different similarities between the embeddings of the clear image and negatives, the learning difficulty of the multiple components is intrinsically imbalanced. To tackle this issue, we customize a curriculum learning strategy to reweight the importance of different negatives. In addition, to improve the interpretability in the feature space, we build a physics-aware dual-branch unit according to the atmospheric scattering model. With the unit, as well as curricular contrastive regularization, we establish our dehazing network, named C2PNet. Extensive experiments demonstrate that our C2PNet significantly outperforms state-of-the-art methods, with extreme PSNR boosts of 3.94dB and 1.50dB, respectively, on SOTS-indoor and SOTS-outdoor datasets. Code is available at https://github.com/YuZheng9/C2PNet.
****
Learning From Noisy Labels With Decoupled Meta Label Purifier
Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Yabiao Wang, Chengjie Wang, Cai Rong Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19934-19943
Training deep neural networks (DNN) with noisy labels is challenging since DNN can easily memorize inaccurate labels, leading to poor generalization ability. Recently, the meta-learning based label correction strategy is widely adopted to tackle this problem via identifying and correcting potential noisy labels with the help of a small set of clean validation data. Although training with purified labels can effectively improve performance, solving the meta-learning problem inevitably involves a nested loop of bi-level optimization between model weights and hyperparameters (i.e., label distribution). As compromise, previous methods resort toa coupled learning process with alternating update. In this paper, we empirically find such simultaneous optimization over both model weights and label distribution can not achieve an optimal routine, consequently limiting the representation ability of backbone and accuracy of corrected labels. From this observation, a novel multi-stage label purifier named DMLP is proposed. DMLP decouples the label correction process into label-free representation learning and a simple meta label purifier, In this way, DMLP can focus on extracting discriminative feature and label correction in two distinctive stages. DMLP is a plug-and-play label purifier, the purified labels can be directly reused in naive end-to-end network retraining or other robust learning methods, where state-of-the-art results are obtained on several synthetic and real-world noisy datasets, especially under high noise levels.
****
Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification
Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, Mark Yatskar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19187-19197
Concept Bottleneck Models (CBM) are inherently interpretable models that factor model decisions into human-readable concepts. They allow people to easily understand why a model is failing, a critical feature for high-stakes applications. CBMs require manually specified concepts and often under-perform their black box counterparts, preventing their broad adoption. We address these shortcomings and are first to show how to construct high-performance CBMs without manual specification of similar accuracy to black box models. Our approach, Language Guided Bottlenecks (LaBo), leverages a language model, GPT-3, to define a large space of p

ossible bottlenecks. Given a problem domain, LaBo uses GPT-3 to produce factual sentences about categories to form candidate concepts. LaBo efficiently searches possible bottlenecks through a novel submodular utility that promotes the selection of discriminative and diverse information. Ultimately, GPT-3's sentential concepts can be aligned to images using CLIP, to form a bottleneck layer. Experiments demonstrate that LaBo is a highly effective prior for concepts important to visual recognition. In the evaluation with 11 diverse datasets, LaBo bottlenecks excel at few-shot classification: they are 11.7% more accurate than black box linear probes at 1 shot and comparable with more data. Overall, LaBo demonstrates that inherently interpretable models can be widely applied at similar, or better, performance than black box approaches.
****

Sharpness-Aware Gradient Matching for Domain Generalization

Pengfei Wang, Zhaoxiang Zhang, Zhen Lei, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3769-3778

The goal of domain generalization (DG) is to enhance the generalization capability of the model learned from a source domain to other unseen domains. The recently developed Sharpness-Aware Minimization (SAM) method aims to achieve this goal by minimizing the sharpness measure of the loss landscape. Though SAM and its variants have demonstrated impressive DG performance, they may not always converge to the desired flat region with a small loss value. In this paper, we present two conditions to ensure that the model could converge to a flat minimum with a small loss, and present an algorithm, named Sharpness-Aware Gradient Matching (SAGM), to meet the two conditions for improving model generalization capability. Specifically, the optimization objective of SAGM will simultaneously minimize the empirical risk, the perturbed loss (i.e., the maximum loss within a neighborhood in the parameter space), and the gap between them. By implicitly aligning the gradient directions between the empirical risk and the perturbed loss, SAGM improves the generalization capability over SAM and its variants without increasing the computational cost. Extensive experimental results show that our proposed SAGM method consistently outperforms the state-of-the-art methods on five DG benchmarks, including PACS, VLCS, OfficeHome, TerraIncognita, and DomainNet. Codes are available at https://github.com/Wang-pengfei/SAGM.
****

ViPLO: Vision Transformer Based Pose-Conditioned Self-Loop Graph for Human-Object Interaction Detection

Jeeseung Park, Jin-Woo Park, Jong-Seok Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17152-17162

Human-Object Interaction (HOI) detection, which localizes and infers relationships between human and objects, plays an important role in scene understanding. Although two-stage HOI detectors have advantages of high efficiency in training and inference, they suffer from lower performance than one-stage methods due to the old backbone networks and the lack of considerations for the HOI perception process of humans in the interaction classifiers. In this paper, we propose Vision Transformer based Pose-Conditioned Self-Loop Graph (ViPLO) to resolve these problems. First, we propose a novel feature extraction method suitable for the Vision Transformer backbone, called masking with overlapped area (MOA) module. The MOA module utilizes the overlapped area between each patch and the given region in the attention function, which addresses the quantization problem when using the Vision Transformer backbone. In addition, we design a graph with a pose-conditioned self-loop structure, which updates the human node encoding with local features of human joints. This allows the classifier to focus on specific human joints to effectively identify the type of interaction, which is motivated by the human perception process for HOI. As a result, ViPLO achieves the state-of-the-art results on two public benchmarks, especially obtaining a +2.07 mAP performance gain on the HICO-DET dataset.
****

Improving Table Structure Recognition With Visual-Alignment Sequential Coordinate Modeling

Yongshuai Huang, Ning Lu, Dapeng Chen, Yibo Li, Zecheng Xie, Shenggao Zhu, Liang cai Gao, Wei Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11134-11143

Table structure recognition aims to extract the logical and physical structure of unstructured table images into a machine-readable format. The latest end-to-end image-to-text approaches simultaneously predict the two structures by two decoders, where the prediction of the physical structure (the bounding boxes of the cells) is based on the representation of the logical structure. However, as the logical representation lacks the local visual information, the previous methods often produce imprecise bounding boxes. To address this issue, we propose an end-to-end sequential modeling framework for table structure recognition called VAST. It contains a novel coordinate sequence decoder triggered by the representation of the non-empty cell from the logical structure decoder. In the coordinate sequence decoder, we model the bounding box coordinates as a language sequence, where the left, top, right and bottom coordinates are decoded sequentially to leverage the inter-coordinate dependency. Furthermore, we propose an auxiliary visual-alignment loss to enforce the logical representation of the non-empty cells to contain more local visual details, which helps produce better cell bounding boxes. Extensive experiments demonstrate that our proposed method can achieve state-of-the-art results in both logical and physical structure recognition. The ablation study also validates that the proposed coordinate sequence decoder and the visual-alignment loss are the keys to the success of our method.

****

MSINet: Twins Contrastive Search of Multi-Scale Interaction for Object ReID

Jianyang Gu, Kai Wang, Hao Luo, Chen Chen, Wei Jiang, Yuqiang Fang, Shanghang Zhang, Yang You, Jian Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19243-19253

Neural Architecture Search (NAS) has been increasingly appealing to the society of object Re-Identification (ReID), for that task-specific architectures significantly improve the retrieval performance. Previous works explore new optimizing targets and search spaces for NAS ReID, yet they neglect the difference of training schemes between image classification and ReID. In this work, we propose a novel Twins Contrastive Mechanism (TCM) to provide more appropriate supervision for ReID architecture search. TCM reduces the category overlaps between the training and validation data, and assists NAS in simulating real-world ReID training schemes. We then design a Multi-Scale Interaction (MSI) search space to search for rational interaction operations between multi-scale features. In addition, we introduce a Spatial Alignment Module (SAM) to further enhance the attention consistency confronted with images from different sources. Under the proposed NAS scheme, a specific architecture is automatically searched, named as MSINet. Extensive experiments demonstrate that our method surpasses state-of-the-art ReID methods on both in-domain and cross-domain scenarios.

****

WIRE: Wavelet Implicit Neural Representations

Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, Richard G. Baraniuk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18507-18516

Implicit neural representations (INRs) have recently advanced numerous vision-related areas. INR performance depends strongly on the choice of activation function employed in its MLP network. A wide range of nonlinearities have been explored, but, unfortunately, current INRs designed to have high accuracy also suffer from poor robustness (to signal noise, parameter variation, etc.). Inspired by harmonic analysis, we develop a new, highly accurate and robust INR that does not exhibit this tradeoff. Our Wavelet Implicit neural REpresentation (WIRE) uses as its activation function the complex Gabor wavelet that is well-known to be optimally concentrated in space--frequency and to have excellent biases for representing images. A wide range of experiments (image denoising, image inpainting, super-resolution, computed tomography reconstruction, image overfitting, and novel view synthesis with neural radiance fields) demonstrate that WIRE defines the new state of the art in INR accuracy, training time, and robustness.

****
Bi-Directional Feature Fusion Generative Adversarial Network for Ultra-High Reso lution Pathological Image Virtual Re-Staining
Kexin Sun, Zhineng Chen, Gongwei Wang, Jun Liu, Xiongjun Ye, Yu-Gang Jiang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2023, pp. 3904-3913
The cost of pathological examination makes virtual re-staining of pathological i mages meaningful. However, due to the ultra-high resolution of pathological imag es, traditional virtual re-staining methods have to divide a WSI image into patc hes for model training and inference. Such a limitation leads to the lack of glo bal information, resulting in observable differences in color, brightness and co ntrast when the re-stained patches are merged to generate an image of larger siz e. We summarize this issue as the square effect. Some existing methods try to so lve this issue through overlapping between patches or simple post-processing. Bu t the former one is not that effective, while the latter one requires carefully tuning. In order to eliminate the square effect, we design a bi-directional feat ure fusion generative adversarial network (BFF-GAN) with a global branch and a l ocal branch. It learns the inter-patch connections through the fusion of global and local features plus patch-wise attention. We perform experiments on both the private dataset RCC and the public dataset ANHIR. The results show that our mod el achieves competitive performance and is able to generate extremely real image s that are deceptive even for experienced pathologists, which means it is of gre at clinical significance.
****
HumanGen: Generating Human Radiance Fields With Explicit Priors
Suyi Jiang, Haoran Jiang, Ziyu Wang, Haimin Luo, Wenzheng Chen, Lan Xu; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023, pp. 12543-12554
Recent years have witnessed the tremendous progress of 3D GANs for generating vi ew-consistent radiance fields with photo-realism. Yet, high-quality generation o f human radiance fields remains challenging, partially due to the limited human- related priors adopted in existing methods. We present HumanGen, a novel 3D huma n generation scheme with detailed geometry and 360deg realistic free-view render ing. It explicitly marries the 3D human generation with various priors from the 2D generator and 3D reconstructor of humans through the design of "anchor image" . We introduce a hybrid feature representation using the anchor image to bridge the latent space of HumanGen with the existing 2D generator. We then adopt a pro nged design to disentangle the generation of geometry and appearance. With the a id of the anchor image, we adapt a 3D reconstructor for fine-grained details syn thesis and propose a two-stage blending scheme to boost appearance generation. E xtensive experiments demonstrate our effectiveness for state-of-the-art 3D human generation regarding geometry details, texture quality, and free-view performan ce. Notably, HumanGen can also incorporate various off-the-shelf 2D latent editi ng methods, seamlessly lifting them into 3D.
****
Bringing Inputs to Shared Domains for 3D Interacting Hands Recovery in the Wild
Gyeongsik Moon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2023, pp. 17028-17037
Despite recent achievements, existing 3D interacting hands recovery methods have shown results mainly on motion capture (MoCap) environments, not on in-the-wild (ITW) ones. This is because collecting 3D interacting hands data in the wild is extremely challenging, even for the 2D data. We present InterWild, which brings MoCap and ITW samples to shared domains for robust 3D interacting hands recover y in the wild with a limited amount of ITW 2D/3D interacting hands data. 3D inte racting hands recovery consists of two sub-problems: 1) 3D recovery of each hand and 2) 3D relative translation recovery between two hands. For the first sub-pr oblem, we bring MoCap and ITW samples to a shared 2D scale space. Although ITW d atasets provide a limited amount of 2D/3D interacting hands, they contain large- scale 2D single hand data. Motivated by this, we use a single hand image as an i nput for the first sub-problem regardless of whether two hands are interacting.

Hence, interacting hands of MoCap datasets are brought to the 2D scale space of single hands of ITW datasets. For the second sub-problem, we bring MoCap and ITW samples to a shared appearance-invariant space. Unlike the first sub-problem, 2D labels of ITW datasets are not helpful for the second sub-problem due to the 3D translation's ambiguity. Hence, instead of relying on ITW samples, we amplify the generalizability of MoCap samples by taking only a geometric feature without an image as an input for the second sub-problem. As the geometric feature is invariant to appearances, MoCap and ITW samples do not suffer from a huge appearance gap between the two datasets. The code is available in https://github.com/facebookresearch/InterWild.
****

Local Connectivity-Based Density Estimation for Face Clustering

Junho Shin, Hyo-Jun Lee, Hyunseop Kim, Jong-Hyeon Baek, Daehyun Kim, Yeong Jun Koh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13621-13629

Recent graph-based face clustering methods predict the connectivity of enormous edges, including false positive edges that link nodes with different classes. However, those false positive edges, which connect negative node pairs, have the risk of integration of different clusters when their connectivity is incorrectly estimated. This paper proposes a novel face clustering method to address this problem. The proposed clustering method employs density-based clustering, which maintains edges that have higher density. For this purpose, we propose a reliable density estimation algorithm based on local connectivity between K nearest neighbors (KNN). We effectively exclude negative pairs from the KNN graph based on the reliable density while maintaining sufficient positive pairs. Furthermore, we develop a pairwise connectivity estimation network to predict the connectivity of the selected edges. Experimental results demonstrate that the proposed clustering method significantly outperforms the state-of-the-art clustering methods on large-scale face clustering datasets and fashion image clustering datasets. Our code is available at https://github.com/illian01/LCE-PCENet
****

Adaptive Zone-Aware Hierarchical Planner for Vision-Language Navigation

Chen Gao, Xingyu Peng, Mi Yan, He Wang, Lirong Yang, Haibing Ren, Hongsheng Li, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14911-14920

The task of Vision-Language Navigation (VLN) is for an embodied agent to reach the global goal according to the instruction. Essentially, during navigation, a series of sub-goals need to be adaptively set and achieved, which is naturally a hierarchical navigation process. However, previous methods leverage a single-step planning scheme, i.e., directly performing navigation action at each step, which is unsuitable for such a hierarchical navigation process. In this paper, we propose an Adaptive Zone-aware Hierarchical Planner (AZHP) to explicitly divides the navigation process into two heterogeneous phases, i.e., sub-goal setting via zone partition/selection (high-level action) and sub-goal executing (low-level action), for hierarchical planning. Specifically, AZHP asynchronously performs two levels of action via the designed State-Switcher Module (SSM). For high-level action, we devise a Scene-aware adaptive Zone Partition (SZP) method to adaptively divide the whole navigation area into different zones on-the-fly. Then the Goal-oriented Zone Selection (GZS) method is proposed to select a proper zone for the current sub-goal. For low-level action, the agent conducts navigation-decision multi-steps in the selected zone. Moreover, we design a Hierarchical RL (HRL) strategy and auxiliary losses with curriculum learning to train the AZHP framework, which provides effective supervision signals for each stage. Extensive experiments demonstrate the superiority of our proposed method, which achieves state-of-the-art performance on three VLN benchmarks (REVERIE, SOON, R2R).
****

Towards Practical Plug-and-Play Diffusion Models

Hyojun Go, Yunsung Lee, Jin-Young Kim, Seunghyun Lee, Myeongho Jeong, Hyun Seung Lee, Seungtaek Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1962-1971

Diffusion-based generative models have achieved remarkable success in image gene
ration. Their guidance formulation allows an external model to plug-and-play con
trol the generation process for various tasks without fine-tuning the diffusion
model. However, the direct use of publicly available off-the-shelf models for gu
idance fails due to their poor performance on noisy inputs. For that, the existi
ng practice is to fine-tune the guidance models with labeled data corrupted with
 noises. In this paper, we argue that this practice has limitations in two aspec
ts: (1) performing on inputs with extremely various noises is too hard for a sin
gle guidance model; (2) collecting labeled datasets hinders scaling up for vario
us tasks. To tackle the limitations, we propose a novel strategy that leverages
multiple experts where each expert is specialized in a particular noise range an
d guides the reverse process of the diffusion at its corresponding timesteps. Ho
wever, as it is infeasible to manage multiple networks and utilize labeled data,
 we present a practical guidance framework termed Practical Plug-And-Play (PPAP)
, which leverages parameter-efficient fine-tuning and data-free knowledge transf
er. We exhaustively conduct ImageNet class conditional generation experiments to
 show that our method can successfully guide diffusion with small trainable para
meters and no labeled data. Finally, we show that image classifiers, depth estim
ators, and semantic segmentation models can guide publicly available GLIDE throu
gh our framework in a plug-and-play manner. Our code is available at https://git
hub.com/riiid/PPAP.
****
Memory-Friendly Scalable Super-Resolution via Rewinding Lottery Ticket Hypothesi
s
Jin Lin, Xiaotong Luo, Ming Hong, Yanyun Qu, Yuan Xie, Zongze Wu; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 14398-14407
Scalable deep Super-Resolution (SR) models are increasingly in demand, whose mem
ory can be customized and tuned to the computational recourse of the platform. T
he existing dynamic scalable SR methods are not memory-friendly enough because m
ulti-scale models have to be saved with a fixed size for each model. Inspired by
 the success of Lottery Tickets Hypothesis (LTH) on image classification, we exp
lore the existence of unstructured scalable SR deep models, that is, we find gra
dual shrinkage sub-networks of extreme sparsity named winning tickets. In this p
aper, we propose a Memory-friendly Scalable SR framework (MSSR). The advantage i
s that only a single scalable model covers multiple SR models with different siz
es, instead of reloading SR models of different sizes. Concretely, MSSR consists
 of the forward and backward stages, the former for model compression and the la
tter for model expansion. In the forward stage, we take advantage of LTH with re
winding weights to progressively shrink the SR model and the pruning-out masks t
hat form nested sets. Moreover, stochastic self-distillation (SSD) is conducted
to boost the performance of sub-networks. By stochastically selecting multiple d
epths, the current model inputs the selected features into the corresponding par
ts in the larger model and improves the performance of the current model based o
n the feedback results of the larger model. In the backward stage, the smaller S
R model could be expanded by recovering and fine-tuning the pruned parameters ac
cording to the pruning-out masks obtained in the forward. Extensive experiments
show the effectiveness of MMSR. The smallest-scale sub-network could achieve the
 sparsity of 94% and outperforms the compared lightweight SR methods.
****
YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object
 Detectors
Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 746
4-7475
Real-time object detection is one of the most important research topics in compu
ter vision. As new approaches regarding architecture optimization and training o
ptimization are continually being developed, we have found two research topics t
hat have spawned when dealing with these latest state-of-the-art methods. To add
ress the topics, we propose a trainable bag-of-freebies oriented solution. We co

mbine the flexible and efficient training tools with the proposed architecture and the compound scaling method. YOLOv7 surpasses all known object detectors in both speed and accuracy in the range from 5 FPS to 120 FPS and has the highest accuracy 56.8% AP among all known realtime object detectors with 30 FPS or higher on GPU V100. Source code is released in https://github.com/ WongKinYiu/yolov7.
****
Deep Deterministic Uncertainty: A New Simple Baseline
Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H.S. Torr, Yarin Gal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24384-24394
Reliable uncertainty from deterministic single-forward pass models is sought after because conventional methods of uncertainty quantification are computationally expensive. We take two complex single-forward-pass uncertainty approaches, DUQ and SNGP, and examine whether they mainly rely on a well-regularized feature space. Crucially, without using their more complex methods for estimating uncertainty, we find that a single softmax neural net with such a regularized feature-space, achieved via residual connections and spectral normalization, outperforms DUQ and SNGP's epistemic uncertainty predictions using simple Gaussian Discriminant Analysis post-training as a separate feature-space density estimator---without fine-tuning on OoD data, feature ensembling, or input pre-procressing. Our conceptually simple Deep Deterministic Uncertainty (DDU) baseline can also be used to disentangle aleatoric and epistemic uncertainty and performs as well as Deep Ensembles, the state-of-the art for uncertainty prediction, on several OoD benchmarks (CIFAR-10/100 vs SVHN/Tiny-ImageNet, ImageNet vs ImageNet-O), active learning settings across different model architectures, as well as in large scale vision tasks like semantic segmentation, while being computationally cheaper.
****
PartDistillation: Learning Parts From Instance Segmentation
Jang Hyun Cho, Philipp Krähenbühl, Vignesh Ramanathan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7152-7161
We present a scalable framework to learn part segmentation from object instance labels. State-of-the-art instance segmentation models contain a surprising amount of part information. However, much of this information is hidden from plain view. For each object instance, the part information is noisy, inconsistent, and incomplete. PartDistillation transfers the part information of an instance segmentation model into a part segmentation model through self-supervised self-training on a large dataset. The resulting segmentation model is robust, accurate, and generalizes well. We evaluate the model on various part segmentation datasets. Our model outperforms supervised part segmentation in zero-shot generalization performance by a large margin. Our model outperforms when finetuned on target datasets compared to supervised counterpart and other baselines especially in few-shot regime. Finally, our model provides a wider coverage of rare parts when evaluated over 10K object classes. Code is at https://github.com/facebookresearch/PartDistillation.
****
DeltaEdit: Exploring Text-Free Training for Text-Driven Image Manipulation
Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6894-6903
Text-driven image manipulation remains challenging in training or inference flexibility. Conditional generative models depend heavily on expensive annotated training data. Meanwhile, recent frameworks, which leverage pre-trained vision-language models, are limited by either per text-prompt optimization or inference-time hyper-parameters tuning. In this work, we propose a novel framework named DeltaEdit to address these problems. Our key idea is to investigate and identify a space, namely delta image and text space that has well-aligned distribution between CLIP visual feature differences of two images and CLIP textual embedding differences of source and target texts. Based on the CLIP delta space, the DeltaEdit network is designed to map the CLIP visual features differences to the editing

directions of StyleGAN at training phase. Then, in inference phase, DeltaEdit predicts the StyleGAN's editing directions from the differences of the CLIP textual features. In this way, DeltaEdit is trained in a text-free manner. Once trained, it can well generalize to various text prompts for zero-shot inference without bells and whistles. Extensive experiments verify that our method achieves competitive performances with other state-of-the-arts, meanwhile with much better flexibility in both training and inference. Code is available at https://github.com/Yueming6568/DeltaEdit
****
Boosting Video Object Segmentation via Space-Time Correspondence Learning
Yurong Zhang, Liulei Li, Wenguan Wang, Rong Xie, Li Song, Wenjun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2246-2256
Current top-leading solutions for video object segmentation (VOS) typically follow a matching-based regime: for each query frame, the segmentation mask is inferred according to its correspondence to previously processed and the first annotated frames. They simply exploit the supervisory signals from the groundtruth masks for learning mask prediction only, without posing any constraint on the space-time correspondence matching, which, however, is the fundamental building block of such regime. To alleviate this crucial yet commonly ignored issue, we devise a correspondence-aware training framework, which boosts matching-based VOS solutions by explicitly encouraging robust correspondence matching during network learning. Through comprehensively exploring the intrinsic coherence in videos on pixel and object levels, our algorithm reinforces the standard, fully supervised training of mask segmentation with label-free, contrastive correspondence learning. Without neither requiring extra annotation cost during training, nor causing speed delay during deployment, nor incurring architectural modification, our algorithm provides solid performance gains on four widely used benchmarks, i.e., DAVIS2016&2017, and YouTube-VOS2018&2019, on the top of famous matching-based VOS solutions. Our implementation will be released.
****
Towards Realistic Long-Tailed Semi-Supervised Learning: Consistency Is All You Need
Tong Wei, Kai Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3469-3478
While long-tailed semi-supervised learning (LTSSL) has received tremendous attention in many real-world classification problems, existing LTSSL algorithms typically assume that the class distributions of labeled and unlabeled data are almost identical. Those LTSSL algorithms built upon the assumption can severely suffer when the class distributions of labeled and unlabeled data are mismatched since they utilize biased pseudo-labels from the model. To alleviate this issue, we propose a new simple method that can effectively utilize unlabeled data of unknown class distributions by introducing the adaptive consistency regularizer (ACR). ACR realizes the dynamic refinery of pseudo-labels for various distributions in a unified formula by estimating the true class distribution of unlabeled data. Despite its simplicity, we show that ACR achieves state-of-the-art performance on a variety of standard LTSSL benchmarks, e.g., an averaged 10% absolute increase of test accuracy against existing algorithms when the class distributions of labeled and unlabeled data are mismatched. Even when the class distributions are identical, ACR consistently outperforms many sophisticated LTSSL algorithms. We carry out extensive ablation studies to tease apart the factors that are most important to ACR's success. Source code is available at https://github.com/Gank0078/ACR.
****
GAPartNet: Cross-Category Domain-Generalizable Object Perception and Manipulation via Generalizable and Actionable Parts
Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7081-7091
For years, researchers have been devoted to generalizable object perception and

manipulation, where cross-category generalizability is highly desired yet underexplored. In this work, we propose to learn such cross-category skills via Generalizable and Actionable Parts (GAParts). By identifying and defining 9 GAPart classes (lids, handles, etc.) in 27 object categories, we construct a large-scale part-centric interactive dataset, GAPartNet, where we provide rich, part-level annotations (semantics, poses) for 8,489 part instances on 1,166 objects. Based on GAPartNet, we investigate three cross-category tasks: part segmentation, part pose estimation, and part-based object manipulation. Given the significant domain gaps between seen and unseen object categories, we propose a robust 3D segmentation method from the perspective of domain generalization by integrating adversarial learning techniques. Our method outperforms all existing methods by a large margin, no matter on seen or unseen categories. Furthermore, with part segmentation and pose estimation results, we leverage the GAPart pose definition to design part-based manipulation heuristics that can generalize well to unseen object categories in both the simulator and the real world.
****
NeRDi: Single-View NeRF Synthesis With Language-Guided Diffusion As General Image Priors
Congyue Deng, Chiyu “Max” Jiang, Charles R. Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20637-20647
2D-to-3D reconstruction is an ill-posed problem, yet humans are good at solving this problem due to their prior knowledge of the 3D world developed over years. Driven by this observation, we propose NeRDi, a single-view NeRF synthesis framework with general image priors from 2D diffusion models. Formulating single-view reconstruction as an image-conditioned 3D generation problem, we optimize the NeRF representations by minimizing a diffusion loss on its arbitrary view renderings with a pretrained image diffusion model under the input-view constraint. We leverage off-the-shelf vision-language models and introduce a two-section language guidance as conditioning inputs to the diffusion model. This is essentially helpful for improving multiview content coherence as it narrows down the general image prior conditioned on the semantic and visual features of the single-view input image. Additionally, we introduce a geometric loss based on estimated depth maps to regularize the underlying 3D geometry of the NeRF. Experimental results on the DTU MVS dataset show that our method can synthesize novel views with higher quality even compared to existing methods trained on this dataset. We also demonstrate our generalizability in zero-shot NeRF synthesis for in-the-wild images.
****
Therbligs in Action: Video Understanding Through Motion Primitives
Eadom Dessalene, Michael Maynord, Cornelia Fermüller, Yiannis Aloimonos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10618-10626
In this paper we introduce a rule-based, compositional, and hierarchical modeling of action using Therbligs as our atoms. Introducing these atoms provides us with a consistent, expressive, contact-centered representation of action. Over the atoms we introduce a differentiable method of rule-based reasoning to regularize for logical consistency. Our approach is complementary to other approaches in that the Therblig-based representations produced by our architecture augment rather than replace existing architectures' representations. We release the first Therblig-centered annotations over two popular video datasets - EPIC Kitchens 100 and 50-Salads. We also broadly demonstrate benefits to adopting Therblig representations through evaluation on the following tasks: action segmentation, action anticipation, and action recognition - observing an average 10.5%/7.53%/6.5% relative improvement, respectively, over EPIC Kitchens and an average 8.9%/6.63%/4.8% relative improvement, respectively, over 50 Salads. Code and data will be made publicly available.
****
InstantAvatar: Learning Avatars From Monocular Video in 60 Seconds
Tianjian Jiang, Xu Chen, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF C

onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16922-16932
In this paper, we take one step further towards real-world applicability of monocular neural avatar reconstruction by contributing InstantAvatar, a system that can reconstruct human avatars from a monocular video within seconds, and these avatars can be animated and rendered at an interactive rate. To achieve this efficiency we propose a carefully designed and engineered system, that leverages emerging acceleration structures for neural fields, in combination with an efficient empty-space skipping strategy for dynamic scenes. We also contribute an efficient implementation that we will make available for research purposes. Compared to existing methods, InstantAvatar converges 130x faster and can be trained in minutes instead of hours. It achieves comparable or even better reconstruction quality and novel pose synthesis results. When given the same time budget, our method significantly outperforms SoTA methods. InstantAvatar can yield acceptable visual quality in as little as 10 seconds training time. For code and more demo results, please refer to https://ait.ethz.ch/InstantAvatar.
****
You Only Segment Once: Towards Real-Time Panoptic Segmentation
Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, Liujuan Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17819-17829
In this paper, we propose YOSO, a real-time panoptic segmentation framework. YOSO predicts masks via dynamic convolutions between panoptic kernels and image feature maps, in which you only need to segment once for both instance and semantic segmentation tasks. To reduce the computational overhead, we design a feature pyramid aggregator for the feature map extraction, and a separable dynamic decoder for the panoptic kernel generation. The aggregator re-parameterizes interpolation-first modules in a convolution-first way, which significantly speeds up the pipeline without any additional costs. The decoder performs multi-head cross-attention via separable dynamic convolution for better efficiency and accuracy. To the best of our knowledge, YOSO is the first real-time panoptic segmentation framework that delivers competitive performance compared to state-of-the-art models. Specifically, YOSO achieves 46.4 PQ, 45.6 FPS on COCO; 52.5 PQ, 22.6 FPS on Cityscapes; 38.0 PQ, 35.4 FPS on ADE20K; and 34.1 PQ, 7.1 FPS on Mapillary Vistas. Code is available at https://github.com/hujiecpp/YOSO.
****
Robust Single Image Reflection Removal Against Adversarial Attacks
Zhenbo Song, Zhenyuan Zhang, Kaihao Zhang, Wenhan Luo, Zhaoxin Fan, Wenqi Ren, Jianfeng Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24688-24698
This paper addresses the problem of robust deep single-image reflection removal (SIRR) against adversarial attacks. Current deep learning based SIRR methods have shown significant performance degradation due to unnoticeable distortions and perturbations on input images. For a comprehensive robustness study, we first conduct diverse adversarial attacks specifically for the SIRR problem, i.e. towards different attacking targets and regions. Then we propose a robust SIRR model, which integrates the cross-scale attention module, the multi-scale fusion module, and the adversarial image discriminator. By exploiting the multi-scale mechanism, the model narrows the gap between features from clean and adversarial images. The image discriminator adaptively distinguishes clean or noisy inputs, and thus further gains reliable robustness. Extensive experiments on Nature, SIR^2, and Real datasets demonstrate that our model remarkably improves the robustness of SIRR across disparate scenes.
****
OmniObject3D: Large-Vocabulary 3D Object Dataset for Realistic Perception, Reconstruction and Generation
Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 803-814
Recent advances in modeling 3D objects mostly rely on synthetic datasets due to

the lack of large-scale real-scanned 3D databases. To facilitate the development of 3D perception, reconstruction, and generation in the real world, we propose OmniObject3D, a large vocabulary 3D object dataset with massive high-quality real-scanned 3D objects. OmniObject3D has several appealing properties: 1) Large Vocabulary: It comprises 6,000 scanned objects in 190 daily categories, sharing common classes with popular 2D datasets (e.g., ImageNet and LVIS), benefiting the pursuit of generalizable 3D representations. 2) Rich Annotations: Each 3D object is captured with both 2D and 3D sensors, providing textured meshes, point clouds, multiview rendered images, and multiple real-captured videos. 3) Realistic Scans: The professional scanners support high-quality object scans with precise shapes and realistic appearances. With the vast exploration space offered by OmniObject3D, we carefully set up four evaluation tracks: a) robust 3D perception, b) novel-view synthesis, c) neural surface reconstruction, and d) 3D object generation. Extensive studies are performed on these four benchmarks, revealing new observations, challenges, and opportunities for future research in realistic 3D vision.
****
PartMix: Regularization Strategy To Learn Part Discovery for Visible-Infrared Person Re-Identification
Minsu Kim, Seungryong Kim, Jungin Park, Seongheon Park, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18621-18632
Modern data augmentation using a mixture-based technique can regularize the models from overfitting to the training data in various computer vision applications, but a proper data augmentation technique tailored for the part-based Visible-Infrared person Re-IDentification (VI-ReID) models remains unexplored. In this paper, we present a novel data augmentation technique, dubbed PartMix, that synthesizes the augmented samples by mixing the part descriptors across the modalities to improve the performance of part-based VI-ReID models. Especially, we synthesize the positive and negative samples within the same and across different identities and regularize the backbone model through contrastive learning. In addition, we also present an entropy-based mining strategy to weaken the adverse impact of unreliable positive and negative samples. When incorporated into existing part-based VI-ReID model, PartMix consistently boosts the performance. We conduct experiments to demonstrate the effectiveness of our PartMix over the existing VI-ReID methods and provide ablation studies.
****
Uncovering the Disentanglement Capability in Text-to-Image Diffusion Models
Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, Shiyu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1900-1910
Generative models have been widely studied in computer vision. Recently, diffusion models have drawn substantial attention due to the high quality of their generated images. A key desired property of image generative models is the ability to disentangle different attributes, which should enable modification towards a style without changing the semantic content, and the modification parameters should generalize to different images. Previous studies have found that generative adversarial networks (GANs) are inherently endowed with such disentanglement capability, so they can perform disentangled image editing without re-training or fine-tuning the network. In this work, we explore whether diffusion models are also inherently equipped with such a capability. Our finding is that for stable diffusion models, by partially changing the input text embedding from a neutral description (e.g., "a photo of person") to one with style (e.g., "a photo of person with smile") while fixing all the Gaussian random noises introduced during the denoising process, the generated images can be modified towards the target style without changing the semantic content. Based on this finding, we further propose a simple, light-weight image editing algorithm where the mixing weights of the two text embeddings are optimized for style matching and content preservation. This entire process only involves optimizing over around 50 parameters and does not fine-tune the diffusion model itself. Experiments show that the proposed met

hod can modify a wide range of attributes, with the performance outperforming di
ffusion-model-based image-editing algorithms that require fine-tuning. The optim
ized weights generalize well to different images. Our code is publicly available
 at https://github.com/UCSB-NLP-Chang/DiffusionDisentanglement.
****
Feature Representation Learning With Adaptive Displacement Generation and Transf
ormer Fusion for Micro-Expression Recognition
Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, Huijuan Zh
ao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2023, pp. 22086-22095
Micro-expressions are spontaneous, rapid and subtle facial movements that can ne
ither be forged nor suppressed. They are very important nonverbal communication
clues, but are transient and of low intensity thus difficult to recognize. Recen
tly deep learning based methods have been developed for micro-expression recogni
tion using feature extraction and fusion techniques, however, targeted feature l
earning and efficient feature fusion still lack further study according to micro
-expression characteristics. To address these issues, we propose a novel framewo
rk Feature Representation Learning with adaptive Displacement Generation and Tra
nsformer fusion (FRL-DGT), in which a convolutional Displacement Generation Modu
le (DGM) with self-supervised learning is used to extract dynamic feature target
ed to the subsequent ME recognition task, and a well-designed Transformer fusion
 mechanism composed of the Transformer-based local fusion module, global fusion
module, and full-face fusion module is applied to extract the multi-level inform
ative feature from the output of the DGM for the final micro-expression predicti
on. Extensive experiments with solid leave-one-subject-out (LOSO) evaluation res
ults have strongly demonstrated the superiority of our proposed FRL-DGT to state
-of-the-art methods.
****
ViewNet: A Novel Projection-Based Backbone With View Pooling for Few-Shot Point
Cloud Classification
Jiajing Chen, Minmin Yang, Senem Velipasalar; Proceedings of the IEEE/CVF Confer
ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17652-17660
Although different approaches have been proposed for 3D point cloud-related task
s, few-shot learning (FSL) of 3D point clouds still remains under-explored. In F
SL, unlike traditional supervised learning, the classes of training and test dat
a do not overlap, and a model needs to recognize unseen classes from only a few
samples. Existing FSL methods for 3D point clouds employ point-based models as t
heir backbone. Yet, based on our extensive experiments and analysis, we first sh
ow that using a point-based backbone is not the most suitable FSL approach, sinc
e (i) a large number of points' features are discarded by the max pooling operat
ion used in 3D point-based backbones, decreasing the ability of representing sha
pe information; (ii)point-based backbones are sensitive to occlusion. To address
 these issues, we propose employing a projection- and 2D Convolutional Neural Ne
twork-based backbone, referred to as the ViewNet, for FSL from 3D point clouds.
Our approach first projects a 3D point cloud onto six different views to allevia
te the issue of missing points. Also, to generate more descriptive and distingui
shing features, we propose View Pooling, which combines different projected plan
e combinations into five groups and performs max-pooling on each of them. The ex
periments performed on the ModelNet40, ScanObjectNN and ModelNet40-C datasets, w
ith cross validation, show that our method consistently outperforms the state-of
-the-art baselines. Moreover, compared to traditional image classification backb
ones, such as ResNet, the proposed ViewNet can extract more distinguishing featu
res from multiple views of a point cloud. We also show that ViewNet can be used
as a backbone with different FSL heads and provides improved performance compare
d to traditionally used backbones.
****
EXIF As Language: Learning Cross-Modal Associations Between Images and Camera Me
tadata
Chenhao Zheng, Ayush Shrivastava, Andrew Owens; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6945-6956

We learn a visual representation that captures information about the camera that recorded a given photo. To do this, we train a multimodal embedding between image patches and the EXIF metadata that cameras automatically insert into image files. Our model represents this metadata by simply converting it to text and then processing it with a transformer. The features that we learn significantly outperform other self-supervised and supervised features on downstream image forensics and calibration tasks. In particular, we successfully localize spliced image regions "zero shot" by clustering the visual embeddings for all of the patches within an image.
****
ANetQA: A Large-Scale Benchmark for Fine-Grained Compositional Reasoning Over Untrimmed Videos
Zhou Yu, Lixiang Zheng, Zhou Zhao, Fei Wu, Jianping Fan, Kui Ren, Jun Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23191-23200
Building benchmarks to systemically analyze different capabilities of video question answering (VideoQA) models is challenging yet crucial. Existing benchmarks often use non-compositional simple questions and suffer from language biases, making it difficult to diagnose model weaknesses incisively. A recent benchmark AGQA poses a promising paradigm to generate QA pairs automatically from pre-annotated scene graphs, enabling it to measure diverse reasoning abilities with granular control. However, its questions have limitations in reasoning about the fine-grained semantics in videos as such information is absent in its scene graphs. To this end, we present ANetQA, a large-scale benchmark that supports fine-grained compositional reasoning over the challenging untrimmed videos from ActivityNet. Similar to AGQA, the QA pairs in ANetQA are automatically generated from annotated video scene graphs. The fine-grained properties of ANetQA are reflected in the following: (i) untrimmed videos with fine-grained semantics; (ii) spatio-temporal scene graphs with fine-grained taxonomies; and (iii) diverse questions generated from fine-grained templates. ANetQA attains 1.4 billion unbalanced and 13.4 million balanced QA pairs, which is an order of magnitude larger than AGQA with a similar number of videos. Comprehensive experiments are performed for state-of-the-art methods. The best model achieves 44.5% accuracy while human performance tops out at 84.5%, leaving sufficient room for improvement.
****
SadTalker: Learning Realistic 3D Motion Coefficients for Stylized Audio-Driven Single Image Talking Face Animation
Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, Fei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8652-8661
Generating talking head videos through a face image and a piece of speech audio still contains many challenges. i.e., unnatural head movement, distorted expression, and identity modification. We argue that these issues are mainly caused by learning from the coupled 2D motion fields. On the other hand, explicitly using 3D information also suffers problems of stiff expression and incoherent video. We present SadTalker, which generates 3D motion coefficients (head pose, expression) of the 3DMM from audio and implicitly modulates a novel 3D-aware face render for talking head generation. To learn the realistic motion coefficients, we explicitly model the connections between audio and different types of motion coefficients individually. Precisely, we present ExpNet to learn the accurate facial expression from audio by distilling both coefficients and 3D-rendered faces. As for the head pose, we design PoseVAE via a conditional VAE to synthesize head motion in different styles. Finally, the generated 3D motion coefficients are mapped to the unsupervised 3D keypoints space of the proposed face render to synthesize the final video. We conducted extensive experiments to show the superior of our method in terms of motion and video quality.
****
HAAV: Hierarchical Aggregation of Augmented Views for Image Captioning
Chia-Wen Kuo, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11039-11049

A great deal of progress has been made in image captioning, driven by research into how to encode the image using pre-trained models. This includes visual encodings (e.g. image grid features or detected objects) and more recently textual encodings (e.g. image tags or text descriptions of image regions). As more advanced encodings are available and incorporated, it is natural to ask: how to efficiently and effectively leverage the heterogeneous set of encodings? In this paper, we propose to regard the encodings as augmented views of the input image. The image captioning model encodes each view independently with a shared encoder efficiently, and a contrastive loss is incorporated across the encoded views in a novel way to improve their representation quality and the model's data efficiency. Our proposed hierarchical decoder then adaptively weighs the encoded views according to their effectiveness for caption generation by first aggregating within each view at the token level, and then across views at the view level. We demonstrate significant performance improvements of +5.6% CIDEr on MS-COCO and +12.9% CIDEr on Flickr30k compared to state of the arts,
****
CLAMP: Prompt-Based Contrastive Learning for Connecting Language and Animal Pose
Xu Zhang, Wen Wang, Zhe Chen, Yufei Xu, Jing Zhang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23272-23281
Animal pose estimation is challenging for existing image-based methods because of limited training data and large intra- and inter-species variances. Motivated by the progress of visual-language research, we propose that pre-trained language models (eg, CLIP) can facilitate animal pose estimation by providing rich prior knowledge for describing animal keypoints in text. However, we found that building effective connections between pre-trained language models and visual animal keypoints is non-trivial since the gap between text-based descriptions and keypoint-based visual features about animal pose can be significant. To address this issue, we introduce a novel prompt-based Contrastive learning scheme for connecting Language and AniMal Pose (CLAMP) effectively. The CLAMP attempts to bridge the gap by adapting the text prompts to the animal keypoints during network training. The adaptation is decomposed into spatial-aware and feature-aware processes, and two novel contrastive losses are devised correspondingly. In practice, the CLAMP enables the first cross-modal animal pose estimation paradigm. Experimental results show that our method achieves state-of-the-art performance under the supervised, few-shot, and zero-shot settings, outperforming image-based methods by a large margin. The code is available at https://github.com/xuzhang1199/CLAMP.
****
Standing Between Past and Future: Spatio-Temporal Modeling for Multi-Camera 3D Multi-Object Tracking
Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17928-17938
This work proposes an end-to-end multi-camera 3D multi-object tracking (MOT) framework. It emphasizes spatio-temporal continuity and integrates both past and future reasoning for tracked objects. Thus, we name it "Past-and-Future reasoning for Tracking" (PF-Track). Specifically, our method adapts the "tracking by attention" framework and represents tracked instances coherently over time with object queries. To explicitly use historical cues, our "Past Reasoning" module learns to refine the tracks and enhance the object features by cross-attending to queries from previous frames and other objects. The "Future Reasoning" module digests historical information and predicts robust future trajectories. In the case of long-term occlusions, our method maintains the object positions and enables re-association by integrating motion predictions. On the nuScenes dataset, our method improves AMOTA by a large margin and remarkably reduces ID-Switches by 90% compared to prior approaches, which is an order of magnitude less. The code and models are made available at https://github.com/TRI-ML/PF-Track.
****
Learning Sample Relationship for Exposure Correction

Jie Huang, Feng Zhao, Man Zhou, Jie Xiao, Naishan Zheng, Kaiwen Zheng, Zhiwei Xi ong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2023, pp. 9904-9913
Exposure correction task aims to correct the underexposure and its adverse overe xposure images to the normal exposure in a single network. As well recognized, t he optimization flow is opposite. Despite the great advancement, existing exposu re correction methods are usually trained with a mini-batch of both underexposur e and overexposure mixed samples and have not explored the relationship between them to solve the optimization inconsistency. In this paper, we introduce a new perspective to conjunct their optimization processes by correlating and constrai ning the relationship of correction procedure in a mini-batch. The core designs of our framework consist of two steps: 1) formulating the exposure relationship of samples across the batch dimension via a context-irrelevant pretext task. 2) delivering the above sample relationship design as the regularization term withi n the loss function to promote optimization consistency. The proposed sample rel ationship design as a general term can be easily integrated into existing exposu re correction methods without any computational burden in inference time. Extens ive experiments over multiple representative exposure correction benchmarks demo nstrate consistent performance gains by introducing our sample relationship desi gn.
****
TRACE: 5D Temporal Regression of Avatars With Dynamic Cameras in 3D Environments
Yu Sun, Qian Bao, Wu Liu, Tao Mei, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8856-88 66
Although the estimation of 3D human pose and shape (HPS) is rapidly progressing, current methods still cannot reliably estimate moving humans in global coordina tes, which is critical for many applications. This is particularly challenging w hen the camera is also moving, entangling human and camera motion. To address th ese issues, we adopt a novel 5D representation (space, time, and identity) that enables end-to-end reasoning about people in scenes. Our method, called TRACE, i ntroduces several novel architectural components. Most importantly, it uses two new "maps" to reason about the 3D trajectory of people over time in camera, and world, coordinates. An additional memory unit enables persistent tracking of peo ple even during long occlusions. TRACE is the first one-stage method to jointly recover and track 3D humans in global coordinates from dynamic cameras. By train ing it end-to-end, and using full image information, TRACE achieves state-of-the -art performance on tracking and HPS benchmarks. The code and dataset are releas ed for research purposes.
****
TTA-COPE: Test-Time Adaptation for Category-Level Object Pose Estimation
Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Sh in, Stan Birchfield, In So Kweon, Kuk-Jin Yoon; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21285-21295
Test-time adaptation methods have been gaining attention recently as a practical solution for addressing source-to-target domain gaps by gradually updating the model without requiring labels on the target data. In this paper, we propose a m ethod of test-time adaptation for category-level object pose estimation called T TA-COPE. We design a pose ensemble approach with a self-training loss using pose -aware confidence. Unlike previous unsupervised domain adaptation methods for ca tegory-level object pose estimation, our approach processes the test data in a s equential, online manner, and it does not require access to the source domain at runtime. Extensive experimental results demonstrate that the proposed pose ense mble and the self-training loss improve category-level object pose performance d uring test time under both semi-supervised and unsupervised settings.
****
TrojDiff: Trojan Attacks on Diffusion Models With Diverse Targets
Weixin Chen, Dawn Song, Bo Li; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2023, pp. 4035-4044
Diffusion models have achieved great success in a range of tasks, such as image

synthesis and molecule design. As such successes hinge on large-scale training d
ata collected from diverse sources, the trustworthiness of these collected data
is hard to control or audit. In this work, we aim to explore the vulnerabilities
 of diffusion models under potential training data manipulations and try to answ
er: How hard is it to perform Trojan attacks on well-trained diffusion models? W
hat are the adversarial targets that such Trojan attacks can achieve? To answer
these questions, we propose an effective Trojan attack against diffusion models,
 TrojDiff, which optimizes the Trojan diffusion and generative processes during
training. In particular, we design novel transitions during the Trojan diffusion
 process to diffuse adversarial targets into a biased Gaussian distribution and
propose a new parameterization of the Trojan generative process that leads to an
 effective training objective for the attack. In addition, we consider three typ
es of adversarial targets: the Trojaned diffusion models will always output inst
ances belonging to a certain class from the in-domain distribution (In-D2D attac
k), out-of-domain distribution (Out-D2D-attack), and one specific instance (D2I
attack). We evaluate TrojDiff on CIFAR-10 and CelebA datasets against both DDPM
and DDIM diffusion models. We show that TrojDiff always achieves high attack per
formance under different adversarial targets using different types of triggers,
while the performance in benign environments is preserved. The code is available
 at https://github.com/chenweixin107/TrojDiff.
****
End-to-End 3D Dense Captioning With Vote2Cap-DETR
Sijin Chen, Hongyuan Zhu, Xin Chen, Yinjie Lei, Gang Yu, Tao Chen; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
3, pp. 11124-11133
3D dense captioning aims to generate multiple captions localized with their asso
ciated object regions. Existing methods follow a sophisticated "detect-then-desc
ribe" pipeline equipped with numerous hand-crafted components. However, these ha
nd-crafted components would yield suboptimal performance given cluttered object
spatial and class distributions among different scenes. In this paper, we propos
e a simple-yet-effective transformer framework Vote2Cap-DETR based on recent pop
ular DEtection TRansformer (DETR). Compared with prior arts, our framework has s
everal appealing advantages: 1) Without resorting to numerous hand-crafted compo
nents, our method is based on a full transformer encoder-decoder architecture wi
th a learnable vote query driven object decoder, and a caption decoder that prod
uces the dense captions in a set-prediction manner. 2) In contrast to the two-st
age scheme, our method can perform detection and captioning in one-stage. 3) Wit
hout bells and whistles, extensive experiments on two commonly used datasets, Sc
anRefer and Nr3D, demonstrate that our Vote2Cap-DETR surpasses current state-of-
the-arts by 11.13% and 7.11% in CIDEr@0.5IoU, respectively. Codes will be releas
ed soon.
****
Mitigating Task Interference in Multi-Task Learning via Explicit Task Routing Wi
th Non-Learnable Primitives
Chuntao Ding, Zhichao Lu, Shangguang Wang, Ran Cheng, Vishnu Naresh Boddeti; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 7756-7765
Multi-task learning (MTL) seeks to learn a single model to accomplish multiple t
asks by leveraging shared information among the tasks. Existing MTL models, howe
ver, have been known to suffer from negative interference among tasks. Efforts t
o mitigate task interference have focused on either loss/gradient balancing or i
mplicit parameter partitioning with partial overlaps among the tasks. In this pa
per, we propose ETR-NLP to mitigate task interference through a synergistic comb
ination of non-learnable primitives (NLPs) and explicit task routing (ETR). Our
key idea is to employ non-learnable primitives to extract a diverse set of task-
agnostic features and recombine them into a shared branch common to all tasks an
d explicit task-specific branches reserved for each task. The non-learnable prim
itives and the explicit decoupling of learnable parameters into shared and task-
specific ones afford the flexibility needed for minimizing task interference. We
 evaluate the efficacy of ETR-NLP networks for both image-level classification a

nd pixel-level dense prediction MTL problems. Experimental results indicate that ETR-NLP significantly outperforms state-of-the-art baselines with fewer learnable parameters and similar FLOPs across all datasets. Code is available at this URL.
****
Learned Two-Plane Perspective Prior Based Image Resampling for Efficient Object Detection
Anurag Ghosh, N. Dinesh Reddy, Christoph Mertz, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13364-13373
Real-time efficient perception is critical for autonomous navigation and city scale sensing. Orthogonal to architectural improvements, streaming perception approaches have exploited adaptive sampling improving real-time detection performance. In this work, we propose a learnable geometry-guided prior that incorporates rough geometry of the 3D scene (a ground plane and a plane above) to resample images for efficient object detection. This significantly improves small and far-away object detection performance while also being more efficient both in terms of latency and memory. For autonomous navigation, using the same detector and scale, our approach improves detection rate by +4.1 AP_S or +39% and in real-time performance by +5.3 sAP_S or +63% for small objects over state-of-the-art (SOTA). For fixed traffic cameras, our approach detects small objects at image scales other methods cannot. At the same scale, our approach improves detection of small objects by 195% (+12.5 AP_S) over naive-downsampling and 63% (+4.2 AP_S) over SOTA.
****
Tell Me What Happened: Unifying Text-Guided Video Completion via Multimodal Masked Video Generation
Tsu-Jui Fu, Licheng Yu, Ning Zhang, Cheng-Yang Fu, Jong-Chyi Su, William Yang Wang, Sean Bell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10681-10692
Generating a video given the first several static frames is challenging as it anticipates reasonable future frames with temporal coherence. Besides video prediction, the ability to rewind from the last frame or infilling between the head and tail is also crucial, but they have rarely been explored for video completion. Since there could be different outcomes from the hints of just a few frames, a system that can follow natural language to perform video completion may significantly improve controllability. Inspired by this, we introduce a novel task, text-guided video completion (TVC), which requests the model to generate a video from partial frames guided by an instruction. We then propose Multimodal Masked Video Generation (MMVG) to address this TVC task. During training, MMVG discretizes the video frames into visual tokens and masks most of them to perform video completion from any time point. At inference time, a single MMVG model can address all 3 cases of TVC, including video prediction, rewind, and infilling, by applying corresponding masking conditions. We evaluate MMVG in various video scenarios, including egocentric, animation, and gaming. Extensive experimental results indicate that MMVG is effective in generating high-quality visual appearances with text guidance for TVC.
****
Tracking Through Containers and Occluders in the Wild
Basile Van Hoorick, Pavel Tokmakov, Simon Stent, Jie Li, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13802-13812
Tracking objects with persistence in cluttered and dynamic environments remains a difficult challenge for computer vision systems. In this paper, we introduce TCOW, a new benchmark and model for visual tracking through heavy occlusion and containment. We set up a task where the goal is to, given a video sequence, segment both the projected extent of the target object, as well as the surrounding container or occluder whenever one exists. To study this task, we create a mixture of synthetic and annotated real datasets to support both supervised learning and structured evaluation of model performance under various forms of task variati

on, such as moving or nested containment. We evaluate two recent transformer-based video models and find that while they can be surprisingly capable of tracking targets under certain settings of task variation, there remains a considerable performance gap before we can claim a tracking model to have acquired a true notion of object permanence.
****
Geometry and Uncertainty-Aware 3D Point Cloud Class-Incremental Semantic Segmentation
Yuwei Yang, Munawar Hayat, Zhao Jin, Chao Ren, Yinjie Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21759-21768
Despite the significant recent progress made on 3D point cloud semantic segmentation, the current methods require training data for all classes at once, and are not suitable for real-life scenarios where new categories are being continuously discovered. Substantial memory storage and expensive re-training is required to update the model to sequentially arriving data for new concepts. In this paper, to continually learn new categories using previous knowledge, we introduce class-incremental semantic segmentation of 3D point cloud. Unlike 2D images, 3D point clouds are disordered and unstructured, making it difficult to store and transfer knowledge especially when the previous data is not available. We further face the challenge of semantic shift, where previous/future classes are indiscriminately collapsed and treated as the background in the current step, causing a dramatic performance drop on past classes. We exploit the structure of point cloud and propose two strategies to address these challenges. First, we design a geometry-aware distillation module that transfers point-wise feature associations in terms of their geometric characteristics. To counter forgetting caused by the semantic shift, we further develop an uncertainty-aware pseudo-labelling scheme that eliminates noise in uncertain pseudo-labels by label propagation within a local neighborhood. Our extensive experiments on S3DIS and ScanNet in a class-incremental setting show impressive results comparable to the joint training strategy (upper bound). Code is available at: https://github.com/leolyj/3DPC-CISS
****
Neural Kernel Surface Reconstruction
Jiahui Huang, Zan Gojcic, Matan Atzmon, Or Litany, Sanja Fidler, Francis Williams; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4369-4379
We present a novel method for reconstructing a 3D implicit surface from a large-scale, sparse, and noisy point cloud. Our approach builds upon the recently introduced Neural Kernel Fields (NKF) representation. It enjoys similar generalization capabilities to NKF, while simultaneously addressing its main limitations: (a) We can scale to large scenes through compactly supported kernel functions, which enable the use of memory-efficient sparse linear solvers. (b) We are robust to noise, through a gradient fitting solve. (c) We minimize training requirements, enabling us to learn from any dataset of dense oriented points, and even mix training data consisting of objects and scenes at different scales. Our method is capable of reconstructing millions of points in a few seconds, and handling very large scenes in an out-of-core fashion. We achieve state-of-the-art results on reconstruction benchmarks consisting of single objects (ShapeNet, ABC), indoor scenes (ScanNet, Matterport3D), and outdoor scenes (CARLA, Waymo).
****
Cooperation or Competition: Avoiding Player Domination for Multi-Target Robustness via Adaptive Budgets
Yimu Wang, Dinghuai Zhang, Yihan Wu, Heng Huang, Hongyang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20564-20574
Despite incredible advances, deep learning has been shown to be susceptible to adversarial attacks. Numerous approaches were proposed to train robust networks both empirically and certifiably. However, most of them defend against only a single type of attack, while recent work steps forward at defending against multiple attacks. In this paper, to understand multi-target robustness, we view this pr

oblem as a bargaining game in which different players (adversaries) negotiate to reach an agreement on a joint direction of parameter updating. We identify a phenomenon named player domination in the bargaining game, and show that with this phenomenon, some of the existing max-based approaches such as MAX and MSD do not converge. Based on our theoretical results, we design a novel framework that adjusts the budgets of different adversaries to avoid player domination. Experiments on two benchmarks show that employing the proposed framework to the existing approaches significantly advances multi-target robustness.
****
Decompose, Adjust, Compose: Effective Normalization by Playing With Frequency for Domain Generalization
Sangrok Lee, Jongseong Bae, Ha Young Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11776-11785
Domain generalization (DG) is a principal task to evaluate the robustness of computer vision models. Many previous studies have used normalization for DG. In normalization, statistics and normalized features are regarded as style and content, respectively. However, it has a content variation problem when removing style because the boundary between content and style is unclear. This study addresses this problem from the frequency domain perspective, where amplitude and phase are considered as style and content, respectively. First, we verify the quantitative phase variation of normalization through the mathematical derivation of the Fourier transform formula. Then, based on this, we propose a novel normalization method, PCNorm, which eliminates style only as the preserving content through spectral decomposition. Furthermore, we propose advanced PCNorm variants, CCNorm and SCNorm, which adjust the degrees of variations in content and style, respectively. Thus, they can learn domain-agnostic representations for DG. With the normalization methods, we propose ResNet-variant models, DAC-P and DAC-SC, which are robust to the domain gap. The proposed models outperform other recent DG methods. The DAC-SC achieves an average state-of-the-art performance of 65.6% on five datasets: PACS, VLCS, Office-Home, DomainNet, and TerraIncognita.
****
Multilateral Semantic Relations Modeling for Image Text Retrieval
Zheng Wang, Zhenwei Gao, Kangshuai Guo, Yang Yang, Xiaoming Wang, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2830-2839
Image-text retrieval is a fundamental task to bridge vision and language by exploiting various strategies to fine-grained alignment between regions and words. This is still tough mainly because of one-to-many correspondence, where a set of matches from another modality can be accessed by a random query. While existing solutions to this problem including multi-point mapping, probabilistic distribution, and geometric embedding have made promising progress, one-to-many correspondence is still under-explored. In this work, we develop a Multilateral Semantic Relations Modeling (termed MSRM) for image-text retrieval to capture the one-to-many correspondence between multiple samples and a given query via hypergraph modeling. Specifically, a given query is first mapped as a probabilistic embedding to learn its true semantic distribution based on Mahalanobis distance. Then each candidate instance in a mini-batch is regarded as a hypergraph node with its mean semantics while a Gaussian query is modeled as a hyperedge to capture the semantic correlations beyond the pair between candidate points and the query. Comprehensive experimental results on two widely used datasets demonstrate that our MSRM method can outperform state-of-the-art methods in the settlement of multiple matches while still maintaining the comparable performance of instance-level matching. Our codes and checkpoints will be released soon.
****
Optimization-Inspired Cross-Attention Transformer for Compressive Sensing
Jiechong Song, Chong Mou, Shiqi Wang, Siwei Ma, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6174-6184
By integrating certain optimization solvers with deep neural networks, deep unfolding network (DUN) with good interpretability and high performance has attracte

d growing attention in compressive sensing (CS). However, existing DUNs often im
prove the visual quality at the price of a large number of parameters and have t
he problem of feature information loss during iteration. In this paper, we propo
se an Optimization-inspired Cross-attention Transformer (OCT) module as an itera
tive process, leading to a lightweight OCT-based Unfolding Framework (OCTUF) for
 image CS. Specifically, we design a novel Dual Cross Attention (Dual-CA) sub-mo
dule, which consists of an Inertia-Supplied Cross Attention (ISCA) block and a P
rojection-Guided Cross Attention (PGCA) block. ISCA block introduces multi-chann
el inertia forces and increases the memory effect by a cross attention mechanism
 between adjacent iterations. And, PGCA block achieves an enhanced information i
nteraction, which introduces the inertia force into the gradient descent step th
rough a cross attention block. Extensive CS experiments manifest that our OCTUF
achieves superior performance compared to state-of-the-art methods while trainin
g lower complexity. Codes are available at https://github.com/songjiechong/OCTUF
.
****

Novel Class Discovery for 3D Point Cloud Semantic Segmentation

Luigi Riz, Cristiano Saltori, Elisa Ricci, Fabio Poiesi; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 939
3-9402

Novel class discovery (NCD) for semantic segmentation is the task of learning a
model that can segment unlabelled (novel) classes using only the supervision fro
m labelled (base) classes. This problem has recently been pioneered for 2D image
 data, but no work exists for 3D point cloud data. In fact, the assumptions made
 for 2D are loosely applicable to 3D in this case. This paper is presented to ad
vance the state of the art on point cloud data analysis in four directions. Firs
tly, we address the new problem of NCD for point cloud semantic segmentation. Se
condly, we show that the transposition of the only existing NCD method for 2D se
mantic segmentation to 3D data is suboptimal. Thirdly, we present a new method f
or NCD based on online clustering that exploits uncertainty quantification to pr
oduce prototypes for pseudo-labelling the points of the novel classes. Lastly, w
e introduce a new evaluation protocol to assess the performance of NCD for point
 cloud semantic segmentation. We thoroughly evaluate our method on SemanticKITTI
 and SemanticPOSS datasets, showing that it can significantly outperform the bas
eline. Project page: https://github.com/LuigiRiz/NOPS.
****

CAT: LoCalization and IdentificAtion Cascade Detection Transformer for Open-Worl
d Object Detection

Shuailei Ma, Yuefeng Wang, Ying Wei, Jiaqi Fan, Thomas H. Li, Hongli Liu, Fanbin
g Lv; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2023, pp. 19681-19690

Open-world object detection (OWOD), as a more general and challenging goal, requ
ires the model trained from data on known objects to detect both known and unkno
wn objects and incrementally learn to identify these unknown objects. The existi
ng works which employ standard detection framework and fixed pseudo-labelling me
chanism (PLM) have the following problems: (i) The inclusion of detecting unknow
n objects substantially reduces the model's ability to detect known ones. (ii) T
he PLM does not adequately utilize the priori knowledge of inputs. (iii) The fix
ed selection manner of PLM cannot guarantee that the model is trained in the rig
ht direction. We observe that humans subconsciously prefer to focus on all foreg
round objects and then identify each one in detail, rather than localize and ide
ntify a single object simultaneously, for alleviating the confusion. This motiva
tes us to propose a novel solution called CAT: LoCalization and IdentificAtion C
ascade Detection Transformer which decouples the detection process via the share
d decoder in the cascade decoding way. In the meanwhile, we propose the self-ada
ptive pseudo-labelling mechanism which combines the model-driven with input-driv
en PLM and self-adaptively generates robust pseudo-labels for unknown objects, s
ignificantly improving the ability of CAT to retrieve unknown objects. Comprehen
sive experiments on two benchmark datasets, i.e., MS-COCO and PASCAL VOC, show t
hat our model outperforms the state-of-the-art in terms of all metrics in the ta

sk of OWOD, incremental object detection (IOD) and open-set detection.
****
TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and L ocalization
Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, Luisa Verdoli va; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2023, pp. 20606-20615
In this paper we present TruFor, a forensic framework that can be applied to a l arge variety of image manipulation methods, from classic cheapfakes to more rece nt manipulations based on deep learning. We rely on the extraction of both high-level and low-level traces through a transformer-based fusion architecture that combines the RGB image and a learned noise-sensitive fingerprint. The latter lea rns to embed the artifacts related to the camera internal and external processin g by training only on real data in a self-supervised manner. Forgeries are detec ted as deviations from the expected regular pattern that characterizes each pris tine image. Looking for anomalies makes the approach able to robustly detect a v ariety of local manipulations, ensuring generalization. In addition to a pixel-l evel localization map and a whole-image integrity score, our approach outputs a reliability map that highlights areas where localization predictions may be erro r-prone. This is particularly important in forensic applications in order to red uce false alarms and allow for a large scale analysis. Extensive experiments on several datasets show that our method is able to reliably detect and localize bo th cheapfakes and deepfakes manipulations outperforming state-of-the-art works. Code is publicly available at https://grip-unina.github.io/TruFor/.
****
LANA: A Language-Capable Navigator for Instruction Following and Generation
Xiaohan Wang, Wenguan Wang, Jiayi Shao, Yi Yang; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19048-19058
Recently, visual-language navigation (VLN) -- entailing robot agents to follow n avigation instructions -- has shown great advance. However, existing literature put most emphasis on interpreting instructions into actions, only delivering "du mb" wayfinding agents. In this article, we devise LANA, a language-capable navig ation agent which is able to not only execute human-written navigation commands, but also provide route descriptions to humans. This is achieved by simultaneous ly learning instruction following and generation with only one single model. Mor e specifically, two encoders, respectively for route and language encoding, are built and shared by two decoders, respectively, for action prediction and instru ction generation, so as to exploit cross-task knowledge and capture task-specifi c characteristics. Throughout pretraining and fine-tuning, both instruction foll owing and generation are set as optimization objectives. We empirically verify t hat, compared with recent advanced task-specific solutions, LANA attains better performances on both instruction following and route description, with nearly ha lf complexity. In addition, endowed with language generation capability, LANA ca n explain to humans its behaviors and assist human's wayfinding. This work is ex pected to foster future efforts towards building more trustworthy and socially-i ntelligent navigation robots. Our code will be released.
****
Learning 3D-Aware Image Synthesis With Unknown Pose Distribution
Zifan Shi, Yujun Shen, Yinghao Xu, Sida Peng, Yiyi Liao, Sheng Guo, Qifeng Chen, Dit-Yan Yeung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2023, pp. 13062-13071
Existing methods for 3D-aware image synthesis largely depend on the 3D pose dist ribution pre-estimated on the training set. An inaccurate estimation may mislead the model into learning faulty geometry. This work proposes PoF3D that frees ge nerative radiance fields from the requirements of 3D pose priors. We first equip the generator with an efficient pose learner, which is able to infer a pose fro m a latent code, to approximate the underlying true pose distribution automatica lly. We then assign the discriminator a task to learn pose distribution under th e supervision of the generator and to differentiate real and synthesized images with the predicted pose as the condition. The pose-free generator and the pose-a

ware discriminator are jointly trained in an adversarial manner. Extensive resul
ts on a couple of datasets confirm that the performance of our approach, regardi
ng both image quality and geometry quality, is on par with state of the art. To
our best knowledge, PoF3D demonstrates the feasibility of learning high-quality
3D-aware image synthesis without using 3D pose priors for the first time. Projec
t page can be found at https://vivianszf.github.io/pof3d/.
****
Normalizing Flow Based Feature Synthesis for Outlier-Aware Object Detection
Nishant Kumar, Siniša Šegvi■, Abouzar Eslami, Stefan Gumhold; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp
. 5156-5165
Real-world deployment of reliable object detectors is crucial for applications s
uch as autonomous driving. However, general-purpose object detectors like Faster
 R-CNN are prone to providing overconfident predictions for outlier objects. Rec
ent outlier-aware object detection approaches estimate the density of instance-w
ide features with class-conditional Gaussians and train on synthesized outlier f
eatures from their low-likelihood regions. However, this strategy does not guara
ntee that the synthesized outlier features will have a low likelihood according
to the other class-conditional Gaussians. We propose a novel outlier-aware objec
t detection framework that distinguishes outliers from inlier objects by learnin
g the joint data distribution of all inlier classes with an invertible normalizi
ng flow. The appropriate sampling of the flow model ensures that the synthesized
 outliers have a lower likelihood than inliers of all object classes, thereby mo
deling a better decision boundary between inlier and outlier objects. Our approa
ch significantly outperforms the state-of-the-art for outlier-aware object detec
tion on both image and video datasets.
****
DivClust: Controlling Diversity in Deep Clustering
Ioannis Maniadis Metaxas, Georgios Tzimiropoulos, Ioannis Patras; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 3418-3428
Clustering has been a major research topic in the field of machine learning, one
 to which Deep Learning has recently been applied with significant success. Howe
ver, an aspect of clustering that is not addressed by existing deep clustering m
ethods, is that of efficiently producing multiple, diverse partitionings for a g
iven dataset. This is particularly important, as a diverse set of base clusterin
gs are necessary for consensus clustering, which has been found to produce bette
r and more robust results than relying on a single clustering. To address this g
ap, we propose DivClust, a diversity controlling loss that can be incorporated i
nto existing deep clustering frameworks to produce multiple clusterings with the
 desired degree of diversity. We conduct experiments with multiple datasets and
deep clustering frameworks and show that: a) our method effectively controls div
ersity across frameworks and datasets with very small additional computational c
ost, b) the sets of clusterings learned by DivClust include solutions that signi
ficantly outperform single-clustering baselines, and c) using an off-the-shelf c
onsensus clustering algorithm, DivClust produces consensus clustering solutions
that consistently outperform single-clustering baselines, effectively improving
the performance of the base deep clustering framework.
****
CAPE: Camera View Position Embedding for Multi-View 3D Object Detection
Kaixin Xiong, Shi Gong, Xiaoqing Ye, Xiao Tan, Ji Wan, Errui Ding, Jingdong Wang
, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2023, pp. 21570-21579
In this paper, we address the problem of detecting 3D objects from multi-view im
ages. Current query-based methods rely on global 3D position embeddings (PE) to
learn the geometric correspondence between images and 3D space. We claim that di
rectly interacting 2D image features with global 3D PE could increase the diffic
ulty of learning view transformation due to the variation of camera extrinsics.
Thus we propose a novel method based on CAmera view Position Embedding, called C
APE. We form the 3D position embeddings under the local camera-view coordinate s

ystem instead of the global coordinate system, such that 3D position embedding i
s free of encoding camera extrinsic parameters. Furthermore, we extend our CAPE
to temporal modeling by exploiting the object queries of previous frames and enc
oding the ego motion for boosting 3D object detection. CAPE achieves the state-o
f-the-art performance (61.0% NDS and 52.5% mAP) among all LiDAR-free methods on
standard nuScenes dataset. Codes and models are available.
****
Train-Once-for-All Personalization
Hong-You Chen, Yandong Li, Yin Cui, Mingda Zhang, Wei-Lun Chao, Li Zhang; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 11818-11827
We study the problem of how to train a "personalization-friendly" model such tha
t given only the task descriptions, the model can be adapted to different end-us
ers' needs, e.g., for accurately classifying different subsets of objects. One b
aseline approach is to train a "generic" model for classifying a wide range of o
bjects, followed by class selection. In our experiments, we however found it sub
optimal, perhaps because the model's weights are kept frozen without being perso
nalized. To address this drawback, we propose Train-once-for-All PERsonalization
 (TAPER), a framework that is trained just once and can later customize a model
for different end-users given their task descriptions. TAPER learns a set of "ba
sis" models and a mixer predictor, such that given the task description, the wei
ghts (not the predictions!) of the basis models can be on the fly combined into
a single "personalized" model. Via extensive experiments on multiple recognition
 tasks, we show that TAPER consistently outperforms the baseline methods in achi
eving a higher personalized accuracy. Moreover, we show that TAPER can synthesiz
e a much smaller model to achieve comparable performance to a huge generic model
, making it "deployment-friendly" to resource-limited end devices. Interestingly
, even without end-users' task descriptions, TAPER can still be specialized to t
he deployed context based on its past predictions, making it even more "personal
ization-friendly".
****
Bi-Directional Distribution Alignment for Transductive Zero-Shot Learning
Zhicai Wang, Yanbin Hao, Tingting Mu, Ouxiang Li, Shuo Wang, Xiangnan He; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 19893-19902
It is well-known that zero-shot learning (ZSL) can suffer severely from the prob
lem of domain shift, where the true and learned data distributions for the unsee
n classes do not match. Although transductive ZSL (TZSL) attempts to improve thi
s by allowing the use of unlabelled examples from the unseen classes, there is s
till a high level of distribution shift. We propose a novel TZSL model (named as
 Bi-VAEGAN), which largely improves the shift by a strengthened distribution ali
gnment between the visual and auxiliary spaces. The key proposal of the model de
sign includes (1) a bi-directional distribution alignment, (2) a simple but effe
ctive L_2-norm based feature normalization approach, and (3) a more sophisticate
d unseen class prior estimation approach. In benchmark evaluation using four dat
asets, Bi-VAEGAN achieves the new state of the arts under both the standard and
generalized TZSL settings. Code could be found at https://github.com/Zhicaiwww/B
i-VAEGAN.
****
FlexNeRF: Photorealistic Free-Viewpoint Rendering of Moving Humans From Sparse V
iews
Vinoj Jayasundara, Amit Agrawal, Nicolas Heron, Abhinav Shrivastava, Larry S. Da
vis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2023, pp. 21118-21127
We present FlexNeRF, a method for photorealistic free-viewpoint rendering of hum
ans in motion from monocular videos. Our approach works well with sparse views,
which is a challenging scenario when the subject is exhibiting fast/complex moti
ons. We propose a novel approach which jointly optimizes a canonical time and po
se configuration, with a pose-dependent motion field and pose-independent tempor
al deformations complementing each other. Thanks to our novel temporal and cycli

c consistency constraints along with additional losses on intermediate represent ation such as segmentation, our approach provides high quality outputs as the ob served views become sparser. We empirically demonstrate that our method signific antly outperforms the state-of-the-art on public benchmark datasets as well as a self-captured fashion dataset. The project page is available at: https://flex-n erf.github.io/.
****
DIFu: Depth-Guided Implicit Function for Clothed Human Reconstruction
Dae-Young Song, HeeKyung Lee, Jeongil Seo, Donghyeon Cho; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 87 38-8747
Recently, implicit function (IF)-based methods for clothed human reconstruction using a single image have received a lot of attention. Most existing methods rel y on a 3D embedding branch using volume such as the skinned multi-person linear (SMPL) model, to compensate for the lack of information in a single image. Beyon d the SMPL, which provides skinned parametric human 3D information, in this pape r, we propose a new IF-based method, DIFu, that utilizes a projected depth prior containing textured and non-parametric human 3D information. In particular, DIF u consists of a generator, an occupancy prediction network, and a texture predic tion network. The generator takes an RGB image of the human front-side as input, and hallucinates the human back-side image. After that, depth maps for front/ba ck images are estimated and projected into 3D volume space. Finally, the occupan cy prediction network extracts a pixel-aligned feature and a voxel-aligned featu re through a 2D encoder and a 3D encoder, respectively, and estimates occupancy using these features. Note that voxel-aligned features are obtained from the pro jected depth maps, thus it can contain detailed 3D information such as hair and cloths. Also, colors of each 3D point are also estimated with the texture infere nce branch. The effectiveness of DIFu is demonstrated by comparing to recent IF-based models quantitatively and qualitatively.
****
Towards Better Gradient Consistency for Neural Signed Distance Functions via Lev el Set Alignment
Baorui Ma, Junsheng Zhou, Yu-Shen Liu, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17724-1 7734
Neural signed distance functions (SDFs) have shown remarkable capability in repr esenting geometry with details. However, without signed distance supervision, it is still a challenge to infer SDFs from point clouds or multi-view images using neural networks. In this paper, we claim that gradient consistency in the field , indicated by the parallelism of level sets, is the key factor affecting the in ference accuracy. Hence, we propose a level set alignment loss to evaluate the p arallelism of level sets, which can be minimized to achieve better gradient cons istency. Our novelty lies in that we can align all level sets to the zero level set by constraining gradients at queries and their projections on the zero level set in an adaptive way. Our insight is to propagate the zero level set to every where in the field through consistent gradients to eliminate uncertainty in the field that is caused by the discreteness of 3D point clouds or the lack of obser vations from multi-view images. Our proposed loss is a general term which can be used upon different methods to infer SDFs from 3D point clouds and multi-view i mages. Our numerical and visual comparisons demonstrate that our loss can signif icantly improve the accuracy of SDFs inferred from point clouds or multi-view im ages under various benchmarks. Code and data are available at https://github.com /mabaorui/TowardsBetterGradient.
****
Zero-Shot Everything Sketch-Based Image Retrieval, and in Explainable Style
Fengyin Lin, Mingkang Li, Da Li, Timothy Hospedales, Yi-Zhe Song, Yonggang Qi; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23349-23358
This paper studies the problem of zero-short sketch-based image retrieval (ZS-SB IR), however with two significant differentiators to prior art (i) we tackle all

 variants (inter-category, intra-category, and cross datasets) of ZS-SBIR with j
ust one network ("everything"), and (ii) we would really like to understand how
this sketch-photo matching operates ("explainable"). Our key innovation lies wit
h the realization that such a cross-modal matching problem could be reduced to c
omparisons of groups of key local patches -- akin to the seasoned "bag-of-words"
 paradigm. Just with this change, we are able to achieve both of the aforementio
ned goals, with the added benefit of no longer requiring external semantic knowl
edge. Technically, ours is a transformer-based cross-modal network, with three n
ovel components (i) a self-attention module with a learnable tokenizer to produc
e visual tokens that correspond to the most informative local regions, (ii) a cr
oss-attention module to compute local correspondences between the visual tokens
across two modalities, and finally (iii) a kernel-based relation network to asse
mble local putative matches and produce an overall similarity metric for a sketc
h-photo pair. Experiments show ours indeed delivers superior performances across
 all ZS-SBIR settings. The all important explainable goal is elegantly achieved
by visualizing cross-modal token correspondences, and for the first time, via sk
etch to photo synthesis by universal replacement of all matched photo patches.
****
Graph Representation for Order-Aware Visual Transformation
Yue Qiu, Yanjun Sun, Fumiya Matsuzawa, Kenji Iwata, Hirokatsu Kataoka; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2023, pp. 22793-22802
This paper proposes a new visual reasoning formulation that aims at discovering
changes between image pairs and their temporal orders. Recognizing scene dynamic
s and their chronological orders is a fundamental aspect of human cognition. The
 aforementioned abilities make it possible to follow step-by-step instructions,
reason about and analyze events, recognize abnormal dynamics, and restore scenes
 to their previous states. However, it remains unclear how well current AI syste
ms perform in these capabilities. Although a series of studies have focused on i
dentifying and describing changes from image pairs, they mainly consider those c
hanges that occur synchronously, thus neglecting potential orders within those c
hanges. To address the above issue, we first propose a visual transformation gra
ph structure for conveying order-aware changes. Then, we benchmarked previous me
thods on our newly generated dataset and identified the issues of existing metho
ds for change order recognition. Finally, we show a significant improvement in o
rder-aware change recognition by introducing a new model that explicitly associa
tes different changes and then identifies changes and their orders in a graph re
presentation.
****
StarCraftImage: A Dataset for Prototyping Spatial Reasoning Methods for Multi-Ag
ent Environments
Sean Kulinski, Nicholas R. Waytowich, James Z. Hare, David I. Inouye; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2023, pp. 22004-22013
Spatial reasoning tasks in multi-agent environments such as event prediction, ag
ent type identification, or missing data imputation are important for multiple a
pplications (e.g., autonomous surveillance over sensor networks and subtasks for
 reinforcement learning (RL)). StarCraft II game replays encode intelligent (and
 adversarial) multi-agent behavior and could provide a testbed for these tasks;
however, extracting simple and standardized representations for prototyping thes
e tasks is laborious and hinders reproducibility. In contrast, MNIST and CIFAR10
, despite their extreme simplicity, have enabled rapid prototyping and reproduci
bility of ML methods. Following the simplicity of these datasets, we construct a
 benchmark spatial reasoning dataset based on StarCraft II replays that exhibit
complex multi-agent behaviors, while still being as easy to use as MNIST and CIF
AR10. Specifically, we carefully summarize a window of 255 consecutive game stat
es to create 3.6 million summary images from 60,000 replays, including all relev
ant metadata such as game outcome and player races. We develop three formats of
decreasing complexity: Hyperspectral images that include one channel for every u
nit type (similar to multispectral geospatial images), RGB images that mimic CIF

AR10, and grayscale images that mimic MNIST. We show how this dataset can be use
d for prototyping spatial reasoning methods. All datasets, code for extraction,
and code for dataset loading can be found at https://starcraftdata.davidinouye.c
om/.
****
Quality-Aware Pre-Trained Models for Blind Image Quality Assessment
Kai Zhao, Kun Yuan, Ming Sun, Mading Li, Xing Wen; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22302-223
13
Blind image quality assessment (BIQA) aims to automatically evaluate the perceiv
ed quality of a single image, whose performance has been improved by deep learni
ng-based methods in recent years. However, the paucity of labeled data somewhat
restrains deep learning-based BIQA methods from unleashing their full potential.
 In this paper, we propose to solve the problem by a pretext task customized for
 BIQA in a self-supervised learning manner, which enables learning representatio
ns from orders of magnitude more data. To constrain the learning process, we pro
pose a quality-aware contrastive loss based on a simple assumption: the quality
of patches from a distorted image should be similar, but vary from patches from
the same image with different degradations and patches from different images. Fu
rther, we improve the existing degradation process and form a degradation space
with the size of roughly 2x10^7. After pre-trained on ImageNet using our method,
 models are more sensitive to image quality and perform significantly better on
downstream BIQA tasks. Experimental results show that our method obtains remarka
ble improvements on popular BIQA datasets.
****
Topology-Guided Multi-Class Cell Context Generation for Digital Pathology
Shahira Abousamra, Rajarsi Gupta, Tahsin Kurc, Dimitris Samaras, Joel Saltz, Cha
o Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2023, pp. 3323-3333
In digital pathology, the spatial context of cells is important for cell classif
ication, cancer diagnosis and prognosis. To model such complex cell context, how
ever, is challenging. Cells form different mixtures, lineages, clusters and hole
s. To model such structural patterns in a learnable fashion, we introduce severa
l mathematical tools from spatial statistics and topological data analysis. We i
ncorporate such structural descriptors into a deep generative model as both cond
itional inputs and a differentiable loss. This way, we are able to generate high
 quality multi-class cell layouts for the first time. We show that the topology-
rich cell layouts can be used for data augmentation and improve the performance
of downstream tasks such as cell classification.
****
Bi-LRFusion: Bi-Directional LiDAR-Radar Fusion for 3D Dynamic Object Detection
Yingjie Wang, Jiajun Deng, Yao Li, Jinshui Hu, Cong Liu, Yu Zhang, Jianmin Ji, W
anli Ouyang, Yanyong Zhang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2023, pp. 13394-13403
LiDAR and Radar are two complementary sensing approaches in that LiDAR specializ
es in capturing an object's 3D shape while Radar provides longer detection range
s as well as velocity hints. Though seemingly natural, how to efficiently combin
e them for improved feature representation is still unclear. The main challenge
arises from that Radar data are extremely sparse and lack height information. Th
erefore, directly integrating Radar features into LiDAR-centric detection networ
ks is not optimal. In this work, we introduce a bi-directional LiDAR-Radar fusio
n framework, termed Bi-LRFusion, to tackle the challenges and improve 3D detecti
on for dynamic objects. Technically, Bi-LRFusion involves two steps: first, it e
nriches Radar's local features by learning important details from the LiDAR bran
ch to alleviate the problems caused by the absence of height information and ext
reme sparsity; second, it combines LiDAR features with the enhanced Radar featur
es in a unified bird's-eye-view representation. We conduct extensive experiments
 on nuScenes and ORR datasets, and show that our Bi-LRFusion achieves state-of-t
he-art performance for detecting dynamic objects. Notably, Radar data in these t
wo datasets have different formats, which demonstrates the generalizability of o

ur method. Codes will be published.
****
Adaptive Graph Convolutional Subspace Clustering
Lai Wei, Zhengwei Chen, Jun Yin, Changming Zhu, Rigui Zhou, Jin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6262-6271
Spectral-type subspace clustering algorithms have shown excellent performance in many subspace clustering applications. The existing spectral-type subspace clustering algorithms either focus on designing constraints for the reconstruction coefficient matrix or feature extraction methods for finding latent features of original data samples. In this paper, inspired by graph convolutional networks, we use the graph convolution technique to develop a feature extraction method and a coefficient matrix constraint simultaneously. And the graph-convolutional operator is updated iteratively and adaptively in our proposed algorithm. Hence, we call the proposed method adaptive graph convolutional subspace clustering (AGCSC). We claim that, by using AGCSC, the aggregated feature representation of original data samples is suitable for subspace clustering, and the coefficient matrix could reveal the subspace structure of the original data set more faithfully. Finally, plenty of subspace clustering experiments prove our conclusions and show that AGCSC outperforms some related methods as well as some deep models.
****
LOCATE: Localize and Transfer Object Parts for Weakly Supervised Affordance Grounding
Gen Li, Varun Jampani, Deqing Sun, Laura Sevilla-Lara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10922-10931
Humans excel at acquiring knowledge through observation. For example, we can learn to use new tools by watching demonstrations. This skill is fundamental for intelligent systems to interact with the world. A key step to acquire this skill is to identify what part of the object affords each action, which is called affordance grounding. In this paper, we address this problem and propose a framework called LOCATE that can identify matching object parts across images, to transfer knowledge from images where an object is being used (exocentric images used for learning), to images where the object is inactive (egocentric ones used to test). To this end, we first find interaction areas and extract their feature embeddings. Then we learn to aggregate the embeddings into compact prototypes (human, object part, and background), and select the one representing the object part. Finally, we use the selected prototype to guide affordance grounding. We do this in a weakly supervised manner, learning only from image-level affordance and object labels. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods by a large margin on both seen and unseen objects.
****
Learning Steerable Function for Efficient Image Resampling
Jiacheng Li, Chang Chen, Wei Huang, Zhiqiang Lang, Fenglong Song, Youliang Yan, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5866-5875
Image resampling is a basic technique that is widely employed in daily applications. Existing deep neural networks (DNNs) have made impressive progress in resampling performance. Yet these methods are still not the perfect substitute for interpolation, due to the issues of efficiency and continuous resampling. In this work, we propose a novel method of Learning Resampling Function (termed LeRF), which takes advantage of both the structural priors learned by DNNs and the locally continuous assumption of interpolation methods. Specifically, LeRF assigns spatially-varying steerable resampling functions to input image pixels and learns to predict the hyper-parameters that determine the orientations of these resampling functions with a neural network. To achieve highly efficient inference, we adopt look-up tables (LUTs) to accelerate the inference of the learned neural network. Furthermore, we design a directional ensemble strategy and edge-sensitive indexing patterns to better capture local structures. Extensive experiments show that our method runs as fast as interpolation, generalizes well to arbitrary tr

ansformations, and outperforms interpolation significantly, e.g., up to 3dB PSNR gain over bicubic for x2 upsampling on Manga109.
****
TokenHPE: Learning Orientation Tokens for Efficient Head Pose Estimation via Transformers
Cheng Zhang, Hai Liu, Yongjian Deng, Bochen Xie, Youfu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8897-8906
Head pose estimation (HPE) has been widely used in the fields of human machine interaction, self-driving, and attention estimation. However, existing methods cannot deal with extreme head pose randomness and serious occlusions. To address these challenges, we identify three cues from head images, namely, neighborhood similarities, significant facial changes, and critical minority relationships. To leverage the observed findings, we propose a novel critical minority relationship-aware method based on the Transformer architecture in which the facial part relationships can be learned. Specifically, we design several orientation tokens to explicitly encode the basic orientation regions. Meanwhile, a novel token guide multi-loss function is designed to guide the orientation tokens as they learn the desired regional similarities and relationships. We evaluate the proposed method on three challenging benchmark HPE datasets. Experiments show that our method achieves better performance compared with state-of-the-art methods. Our code is publicly available at https://github.com/zc2023/TokenHPE.
****
BioNet: A Biologically-Inspired Network for Face Recognition
Pengyu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10344-10354
Recently, whether and how cutting-edge Neuroscience findings can inspire Artificial Intelligence (AI) confuse both communities and draw much discussion. As one of the most critical fields in AI, Computer Vision (CV) also pays much attention to the discussion. To show our ideas and experimental evidence to the discussion, we focus on one of the most broadly researched topics both in Neuroscience and CV fields, i.e., Face Recognition (FR). Neuroscience studies show that face attributes are essential to the human face-recognizing system. How the attributes contribute also be explained by the Neuroscience community. Even though a few CV works improved the FR performance with attribute enhancement, none of them are inspired by the human face-recognizing mechanism nor boosted performance significantly. To show our idea experimentally, we model the biological characteristics of the human face-recognizing system with classical Convolutional Neural Network Operators (CNN Ops) purposely. We name the proposed Biologically-inspired Network as BioNet. Our BioNet consists of two cascade sub-networks, i.e., the Visual Cortex Network (VCN) and the Inferotemporal Cortex Network (ICN). The VCN is modeled with a classical CNN backbone. The proposed ICN comprises three biologically-inspired modules, i.e., the Cortex Functional Compartmentalization, the Compartment Response Transform, and the Response Intensity Modulation. The experiments prove that: 1) The cutting-edge findings about the human face-recognizing system can further boost the CNN-based FR network. 2) With the biological mechanism, both identity-related attributes (e.g., gender) and identity-unrelated attributes (e.g., expression) can benefit the deep FR models. Surprisingly, the identity-unrelated ones contribute even more than the identity-related ones. 3) The proposed BioNet significantly boosts state-of-the-art on standard FR benchmark datasets. For example, BioNet boosts IJB-B@1e-6 from 52.12% to 68.28% and MegaFace from 98.74% to 99.19%. The source code will be released.
****
Scaling Up GANs for Text-to-Image Synthesis
Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, Taesung Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10124-10134
The recent success of text-to-image synthesis has taken the world by storm and captured the general public's imagination. From a technical standpoint, it also marked a drastic change in the favored architecture to design generative image mo

dels. GANs used to be the de facto choice, with techniques like StyleGAN. With DALL-E 2, auto-regressive and diffusion models became the new standard for large-scale generative models overnight. This rapid shift raises a fundamental question: can we scale up GANs to benefit from large datasets like LAION? We find that naively increasing the capacity of the StyleGAN architecture quickly becomes unstable. We introduce GigaGAN, a new GAN architecture that far exceeds this limit, demonstrating GANs as a viable option for text-to-image synthesis. GigaGAN offers three major advantages. First, it is orders of magnitude faster at inference time, taking only 0.13 seconds to synthesize a 512px image. Second, it can synthesize high-resolution images, for example, 16-megapixel images in 3.66 seconds. Finally, GigaGAN supports various latent space editing applications such as latent interpolation, style mixing, and vector arithmetic operations.
****
DepGraph: Towards Any Structural Pruning
Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16091-16101
Structural pruning enables model acceleration by removing structurally-grouped parameters from neural networks. However, the parameter-grouping patterns vary widely across different models, making architecture-specific pruners, which rely on manually-designed grouping schemes, non-generalizable to new architectures. In this work, we study a highly-challenging yet barely-explored task, any structural pruning, to tackle general structural pruning of arbitrary architecture like CNNs, RNNs, GNNs and Transformers. The most prominent obstacle towards this goal lies in the structural coupling, which not only forces different layers to be pruned simultaneously, but also expects all removed parameters to be consistently unimportant, thereby avoiding structural issues and significant performance degradation after pruning. To address this problem, we propose a general and fully automatic method, Dependency Graph (DepGraph), to explicitly model the dependency between layers and comprehensively group coupled parameters for pruning. In this work, we extensively evaluate our method on several architectures and tasks, including ResNe(X)t, DenseNet, MobileNet and Vision transformer for images, GAT for graph, DGCNN for 3D point cloud, alongside LSTM for language, and demonstrate that, even with a simple norm-based criterion, the proposed method consistently yields gratifying performances.
****
Exploring Discontinuity for Video Frame Interpolation
Sangjin Lee, Hyeongmin Lee, Chajin Shin, Hanbin Son, Sangyoun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9791-9800
Video frame interpolation (VFI) is the task that synthesizes the intermediate frame given two consecutive frames. Most of the previous studies have focused on appropriate frame warping operations and refinement modules for the warped frames. These studies have been conducted on natural videos containing only continuous motions. However, many practical videos contain various unnatural objects with discontinuous motions such as logos, user interfaces and subtitles. We propose three techniques that can make the existing deep learning-based VFI architectures robust to these elements. First is a novel data augmentation strategy called figure-text mixing (FTM) which can make the models learn discontinuous motions during training stage without any extra dataset. Second, we propose a simple but effective module that predicts a map called discontinuity map (D-map), which densely distinguishes between areas of continuous and discontinuous motions. Lastly, we propose loss functions to give supervisions of the discontinuous motion areas which can be applied along with FTM and D-map. We additionally collect a special test benchmark called Graphical Discontinuous Motion (GDM) dataset consisting of some mobile games and chatting videos. Applied to the various state-of-the-art VFI networks, our method significantly improves the interpolation qualities on the videos from not only GDM dataset, but also the existing benchmarks containing only continuous motions such as Vimeo90K, UCF101, and DAVIS.
****

DynamicStereo: Consistent Dynamic Depth From Stereo Videos
Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, Christian Rupprecht; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13229-13239
We consider the problem of reconstructing a dynamic scene observed from a stereo camera. Most existing methods for depth from stereo treat different stereo frames independently, leading to temporally inconsistent depth predictions. Temporal consistency is especially important for immersive AR or VR scenarios, where flickering greatly diminishes the user experience. We propose DynamicStereo, a novel transformer-based architecture to estimate disparity for stereo videos. The network learns to pool information from neighboring frames to improve the temporal consistency of its predictions. Our architecture is designed to process stereo videos efficiently through divided attention layers. We also introduce Dynamic Replica, a new benchmark dataset containing synthetic videos of people and animals in scanned environments, which provides complementary training and evaluation data for dynamic stereo closer to real applications than existing datasets. Training with this dataset further improves the quality of predictions of our proposed DynamicStereo as well as prior methods. Finally, it acts as a benchmark for consistent stereo methods.
****
Cut and Learn for Unsupervised Object Detection and Instance Segmentation
Xudong Wang, Rohit Girdhar, Stella X. Yu, Ishan Misra; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3124-3134
We propose Cut-and-LEaRn (CutLER), a simple approach for training unsupervised object detection and segmentation models. We leverage the property of self-supervised models to 'discover' objects without supervision and amplify it to train a state-of-the-art localization model without any human labels. CutLER first uses our proposed MaskCut approach to generate coarse masks for multiple objects in an image, and then learns a detector on these masks using our robust loss function. We further improve performance by self-training the model on its predictions. Compared to prior work, CutLER is simpler, compatible with different detection architectures, and detects multiple objects. CutLER is also a zero-shot unsupervised detector and improves detection performance AP_50 by over 2.7x on 11 benchmarks across domains like video frames, paintings, sketches, etc. With finetuning, CutLER serves as a low-shot detector surpassing MoCo-v2 by 7.3% AP^box and 6.6% AP^mask on COCO when training with 5% labels.
****
Privacy-Preserving Adversarial Facial Features
Zhibo Wang, He Wang, Shuaifan Jin, Wenwen Zhang, Jiahui Hu, Yan Wang, Peng Sun, Wei Yuan, Kaixin Liu, Kui Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8212-8221
Face recognition service providers protect face privacy by extracting compact and discriminative facial features (representations) from images, and storing the facial features for real-time recognition. However, such features can still be exploited to recover the appearance of the original face by building a reconstruction network. Although several privacy-preserving methods have been proposed, the enhancement of face privacy protection is at the expense of accuracy degradation. In this paper, we propose an adversarial features-based face privacy protection (AdvFace) approach to generate privacy-preserving adversarial features, which can disrupt the mapping from adversarial features to facial images to defend against reconstruction attacks. To this end, we design a shadow model which simulates the attackers' behavior to capture the mapping function from facial features to images and generate adversarial latent noise to disrupt the mapping. The adversarial features rather than the original features are stored in the server's database to prevent leaked features from exposing facial information. Moreover, the AdvFace requires no changes to the face recognition network and can be implemented as a privacy-enhancing plugin in deployed face recognition systems. Extensive experimental results demonstrate that AdvFace outperforms the state-of-the-art face privacy-preserving methods in defending against reconstruction attacks

while maintaining face recognition accuracy.
****
Exploring the Relationship Between Architectural Design and Adversarially Robust Generalization
Aishan Liu, Shiyu Tang, Siyuan Liang, Ruihao Gong, Boxi Wu, Xianglong Liu, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4096-4107
Adversarial training has been demonstrated to be one of the most effective remedies for defending adversarial examples, yet it often suffers from the huge robustness generalization gap on unseen testing adversaries, deemed as the adversarially robust generalization problem. Despite the preliminary understandings devoted to adversarially robust generalization, little is known from the architectural perspective. To bridge the gap, this paper for the first time systematically investigated the relationship between adversarially robust generalization and architectural design. In particular, we comprehensively evaluated 20 most representative adversarially trained architectures on ImageNette and CIFAR-10 datasets towards multiple l_p-norm adversarial attacks. Based on the extensive experiments, we found that, under aligned settings, Vision Transformers (e.g., PVT, CoAtNet) often yield better adversarially robust generalization while CNNs tend to overfit on specific attacks and fail to generalize on multiple adversaries. To better understand the nature behind it, we conduct theoretical analysis via the lens of Rademacher complexity. We revealed the fact that the higher weight sparsity contributes significantly towards the better adversarially robust generalization of Transformers, which can be often achieved by the specially-designed attention blocks. We hope our paper could help to better understand the mechanism for designing robust DNNs. Our model weights can be found at http://robust.art.
****
Vid2Avatar: 3D Avatar Reconstruction From Videos in the Wild via Self-Supervised Scene Decomposition
Chen Guo, Tianjian Jiang, Xu Chen, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12858-12868
We present Vid2Avatar, a method to learn human avatars from monocular in-the-wild videos. Reconstructing humans that move naturally from monocular in-the-wild videos is difficult. Solving it requires accurately separating humans from arbitrary backgrounds. Moreover, it requires reconstructing detailed 3D surface from short video sequences, making it even more challenging. Despite these challenges, our method does not require any groundtruth supervision or priors extracted from large datasets of clothed human scans, nor do we rely on any external segmentation modules. Instead, it solves the tasks of scene decomposition and surface reconstruction directly in 3D by modeling both the human and the background in the scene jointly, parameterized via two separate neural fields. Specifically, we define a temporally consistent human representation in canonical space and formulate a global optimization over the background model, the canonical human shape and texture, and per-frame human pose parameters. A coarse-to-fine sampling strategy for volume rendering and novel objectives are introduced for a clean separation of dynamic human and static background, yielding detailed and robust 3D human reconstructions. The evaluation of our method shows improvements over prior art on publicly available datasets.
****
Task Residual for Tuning Vision-Language Models
Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10899-10909
Large-scale vision-language models (VLMs) pre-trained on billion-level data have learned general visual representations and broad visual concepts. In principle, the well-learned knowledge structure of the VLMs should be inherited appropriately when being transferred to downstream tasks with limited data. However, most existing efficient transfer learning (ETL) approaches for VLMs either damage or are excessively biased towards the prior knowledge, e.g., prompt tuning (PT) dis

cards the pre-trained text-based classifier and builds a new one while adapter-style tuning (AT) fully relies on the pre-trained features. To address this, we propose a new efficient tuning approach for VLMs named Task Residual Tuning (TaskRes), which performs directly on the text-based classifier and explicitly decouples the prior knowledge of the pre-trained models and new knowledge regarding a target task. Specifically, TaskRes keeps the original classifier weights from the VLMs frozen and obtains a new classifier for the target task by tuning a set of prior-independent parameters as a residual to the original one, which enables reliable prior knowledge preservation and flexible task-specific knowledge exploration. The proposed TaskRes is simple yet effective, which significantly outperforms previous ETL methods (e.g., PT and AT) on 11 benchmark datasets while requiring minimal effort for the implementation. Our code is available at https://github.com/geekyutao/TaskRes.
****
Side Adapter Network for Open-Vocabulary Semantic Segmentation
Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2945-2954
This paper presents a new framework for open-vocabulary semantic segmentation with the pre-trained vision-language model, named SAN. Our approach models the semantic segmentation task as a region recognition problem. A side network is attached to a frozen CLIP model with two branches: one for predicting mask proposals, and the other for predicting attention bias which is applied in the CLIP model to recognize the class of masks. This decoupled design has the benefit CLIP in recognizing the class of mask proposals. Since the attached side network can reuse CLIP features, it can be very light. In addition, the entire network can be trained end-to-end, allowing the side network to be adapted to the frozen CLIP model, which makes the predicted mask proposals CLIP-aware. Our approach is fast, accurate, and only adds a few additional trainable parameters. We evaluate our approach on multiple semantic segmentation benchmarks. Our method significantly outperforms other counterparts, with up to 18 times fewer trainable parameters and 19 times faster inference speed. We hope our approach will serve as a solid baseline and help ease future research in open-vocabulary semantic segmentation.
****
Network Expansion for Practical Training Acceleration
Ning Ding, Yehui Tang, Kai Han, Chao Xu, Yunhe Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20269-20279
Recently, the sizes of deep neural networks and training datasets both increase drastically to pursue better performance in a practical sense. With the prevalence of transformer-based models in vision tasks, even more pressure is laid on the GPU platforms to train these heavy models, which consumes a large amount of time and computing resources as well. Therefore, it's crucial to accelerate the training process of deep neural networks. In this paper, we propose a general network expansion method to reduce the practical time cost of the model training process. Specifically, we utilize both width- and depth-level sparsity of dense models to accelerate the training of deep neural networks. Firstly, we pick a sparse sub-network from the original dense model by reducing the number of parameters as the starting point of training. Then the sparse architecture will gradually expand during the training procedure and finally grow into a dense one. We design different expanding strategies to grow CNNs and ViTs respectively, due to the great heterogeneity in between the two architectures. Our method can be easily integrated into popular deep learning frameworks, which saves considerable training time and hardware resources. Extensive experiments show that our acceleration method can significantly speed up the training process of modern vision models on general GPU devices with negligible performance drop (e.g. 1.42x faster for ResNet-101 and 1.34x faster for DeiT-base on ImageNet-1k). The code is available at https://github.com/huawei-noah/Efficient-Computing/tree/master/TrainingAcceleration/NetworkExpansion and https://gitee.com/mindspore/hub/blob/master/mshub_res/assets/noah-cvlab/gpu/1.8/networkexpansion_v1.0_imagenet2012.md.

****
FCC: Feature Clusters Compression for Long-Tailed Visual Recognition
Jian Li, Ziyao Meng, Daqian Shi, Rui Song, Xiaolei Diao, Jingwen Wang, Hao Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24080-24089
Deep Neural Networks (DNNs) are rather restrictive in long-tailed data, since they commonly exhibit an under-representation for minority classes. Various remedies have been proposed to tackle this problem from different perspectives, but they ignore the impact of the density of Backbone Features (BFs) on this issue. Through representation learning, DNNs can map BFs into dense clusters in feature space, while the features of minority classes often show sparse clusters. In practical applications, these features are discretely mapped or even cross the decision boundary resulting in misclassification. Inspired by this observation, we propose a simple and generic method, namely Feature Clusters Compression (FCC), to increase the density of BFs by compressing backbone feature clusters. The proposed FCC can be easily achieved by only multiplying original BFs by a scaling factor in training phase, which establishes a linear compression relationship between the original and multiplied features, and forces DNNs to map the former into denser clusters. In test phase, we directly feed original features without multiplying the factor to the classifier, such that BFs of test samples are mapped closer together and do not easily cross the decision boundary. Meanwhile, FCC can be friendly combined with existing long-tailed methods and further boost them. We apply FCC to numerous state-of-the-art methods and evaluate them on widely used long-tailed benchmark datasets. Extensive experiments fully verify the effectiveness and generality of our method. Code is available at https://github.com/lijian16/FCC.
****
Rethinking the Learning Paradigm for Dynamic Facial Expression Recognition
Hanyang Wang, Bo Li, Shuang Wu, Siyuan Shen, Feng Liu, Shouhong Ding, Aimin Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17958-17968
Dynamic Facial Expression Recognition (DFER) is a rapidly developing field that focuses on recognizing facial expressions in video format. Previous research has considered non-target frames as noisy frames, but we propose that it should be treated as a weakly supervised problem. We also identify the imbalance of short- and long-term temporal relationships in DFER. Therefore, we introduce the Multi-3D Dynamic Facial Expression Learning (M3DFEL) framework, which utilizes Multi-Instance Learning (MIL) to handle inexact labels. M3DFEL generates 3D-instances to model the strong short-term temporal relationship and utilizes 3DCNNs for feature extraction. The Dynamic Long-term Instance Aggregation Module (DLIAM) is then utilized to learn the long-term temporal relationships and dynamically aggregate the instances. Our experiments on DFEW and FERV39K datasets show that M3DFEL outperforms existing state-of-the-art approaches with a vanilla R3D18 backbone. The source code is available at https://github.com/faceeyes/M3DFEL.
****
Multi-Centroid Task Descriptor for Dynamic Class Incremental Inference
Tenghao Cai, Zhizhong Zhang, Xin Tan, Yanyun Qu, Guannan Jiang, Chengjie Wang, Yuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7298-7307
Incremental learning could be roughly divided into two categories, i.e., class- and task-incremental learning. The main difference is whether the task ID is given during evaluation. In this paper, we show this task information is indeed a strong prior knowledge, which will bring significant improvement over class-incremental learning baseline, e.g., DER. Based on this observation, we propose a gate network to predict the task ID for class incremental inference. This is challenging as there is no explicit semantic relationship between categories in the concept of task. Therefore, we propose a multi-centroid task descriptor by assuming the data within a task can form multiple clusters. The cluster centers are optimized by pulling relevant sample-centroid pairs while pushing others away, which ensures that there is at least one centroid close to a given sample. To select

relevant pairs, we use class prototypes as proxies and solve a bipartite matching problem, making the task descriptor representative yet not degenerate to uni-modal. As a result, our dynamic inference network is trained independently of baseline and provides a flexible, efficient solution to distinguish between tasks. Extensive experiments show our approach achieves state-of-the-art results, e.g., we achieve 72.41% average accuracy on CIFAR100-B0S50, outperforming DER by 3.40%.
****
Hierarchical Prompt Learning for Multi-Task Learning
Yajing Liu, Yuning Lu, Hao Liu, Yaozu An, Zhuoran Xu, Zhuokun Yao, Baofeng Zhang, Zhiwei Xiong, Chenguang Gui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10888-10898
Vision-language models (VLMs) can effectively transfer to various vision tasks via prompt learning. Real-world scenarios often require adapting a model to multiple similar yet distinct tasks. Existing methods focus on learning a specific prompt for each task, limiting the ability to exploit potentially shared information from other tasks. Naively training a task-shared prompt using a combination of all tasks ignores fine-grained task correlations. Significant discrepancies across tasks could cause negative transferring. Considering this, we present Hierarchical Prompt (HiPro) learning, a simple and effective method for jointly adapting a pre-trained VLM to multiple downstream tasks. Our method quantifies inter-task affinity and subsequently constructs a hierarchical task tree. Task-shared prompts learned by internal nodes explore the information within the corresponding task group, while task-individual prompts learned by leaf nodes obtain fine-grained information targeted at each task. The combination of hierarchical prompts provides high-quality content of different granularity. We evaluate HiPro on four multi-task learning datasets. The results demonstrate the effectiveness of our method.
****
Physics-Guided ISO-Dependent Sensor Noise Modeling for Extreme Low-Light Photography
Yue Cao, Ming Liu, Shuai Liu, Xiaotao Wang, Lei Lei, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5744-5753
Although deep neural networks have achieved astonishing performance in many vision tasks, existing learning-based methods are far inferior to the physical model-based solutions in extreme low-light sensor noise modeling. To tap the potential of learning-based sensor noise modeling, we investigate the noise formation in a typical imaging process and propose a novel physics-guided ISO-dependent sensor noise modeling approach. Specifically, we build a normalizing flow-based framework to represent the complex noise characteristics of CMOS camera sensors. Each component of the noise model is dedicated to a particular kind of noise under the guidance of physical models. Moreover, we take into consideration of the ISO dependence in the noise model, which is not completely considered by the existing learning-based methods. For training the proposed noise model, a new dataset is further collected with paired noisy-clean images, as well as flat-field and bias frames covering a wide range of ISO settings. Compared to existing methods, the proposed noise model benefits from the flexible structure and accurate modeling capabilities, which can help achieve better denoising performance in extreme low-light scenes. The source code and collected dataset will be publicly available.
****
RIFormer: Keep Your Vision Backbone Effective but Removing Token Mixer
Jiahao Wang, Songyang Zhang, Yong Liu, Taiqiang Wu, Yujiu Yang, Xihui Liu, Kai Chen, Ping Luo, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14443-14452
This paper studies how to keep a vision backbone effective while removing token mixers in its basic building blocks. Token mixers, as self-attention for vision transformers (ViTs), are intended to perform information communication between different spatial tokens but suffer from considerable computational cost and late

ncy. However, directly removing them will lead to an incomplete model structure prior, and thus brings a significant accuracy drop. To this end, we first develop an RepIdentityFormer base on the re-parameterizing idea, to study the token mixer free model architecture. And we then explore the improved learning paradigm to break the limitation of simple token mixer free backbone, and summarize the empirical practice into 5 guidelines. Equipped with the proposed optimization strategy, we are able to build an extremely simple vision backbone with encouraging performance, while enjoying the high efficiency during inference. Extensive experiments and ablative analysis also demonstrate that the inductive bias of network architecture, can be incorporated into simple network structure with appropriate optimization strategy. We hope this work can serve as a starting point for the exploration of optimization-driven efficient network design.
****
Context-Based Trit-Plane Coding for Progressive Image Compression
Seungmin Jeon, Kwang Pyo Choi, Youngo Park, Chang-Su Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14348-14357
Trit-plane coding enables deep progressive image compression, but it cannot use autoregressive context models. In this paper, we propose the context-based trit-plane coding (CTC) algorithm to achieve progressive compression more compactly. First, we develop the context-based rate reduction module to estimate trit probabilities of latent elements accurately and thus encode the trit-planes compactly. Second, we develop the context-based distortion reduction module to refine partial latent tensors from the trit-planes and improve the reconstructed image quality. Third, we propose a retraining scheme for the decoder to attain better rate-distortion tradeoffs. Extensive experiments show that CTC outperforms the baseline trit-plane codec significantly, e.g. by -14.84% in BD-rate on the Kodak lossless dataset, while increasing the time complexity only marginally. The source codes are available at https://github.com/seungminjeon-github/CTC.
****
Self-Supervised Learning for Multimodal Non-Rigid 3D Shape Matching
Dongliang Cao, Florian Bernard; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17735-17744
The matching of 3D shapes has been extensively studied for shapes represented as surface meshes, as well as for shapes represented as point clouds. While point clouds are a common representation of raw real-world 3D data (e.g. from laser scanners), meshes encode rich and expressive topological information, but their creation typically requires some form of (often manual) curation. In turn, methods that purely rely on point clouds are unable to meet the matching quality of mesh-based methods that utilise the additional topological structure. In this work we close this gap by introducing a self-supervised multimodal learning strategy that combines mesh-based functional map regularisation with a contrastive loss that couples mesh and point cloud data. Our shape matching approach allows to obtain intramodal correspondences for triangle meshes, complete point clouds, and partially observed point clouds, as well as correspondences across these data modalities. We demonstrate that our method achieves state-of-the-art results on several challenging benchmark datasets even in comparison to recent supervised methods, and that our method reaches previously unseen cross-dataset generalisation ability.
****
Recurrent Vision Transformers for Object Detection With Event Cameras
Mathias Gehrig, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13884-13893
We present Recurrent Vision Transformers (RVTs), a novel backbone for object detection with event cameras. Event cameras provide visual information with sub-millisecond latency at a high-dynamic range and with strong robustness against motion blur. These unique properties offer great potential for low-latency object detection and tracking in time-critical scenarios. Prior work in event-based vision has achieved outstanding detection performance but at the cost of substantial inference time, typically beyond 40 milliseconds. By revisiting the high-level d

esign of recurrent vision backbones, we reduce inference time by a factor of 6 while retaining similar performance. To achieve this, we explore a multi-stage design that utilizes three key concepts in each stage: First, a convolutional prior that can be regarded as a conditional positional embedding. Second, local- and dilated global self-attention for spatial feature interaction. Third, recurrent temporal feature aggregation to minimize latency while retaining temporal information. RVTs can be trained from scratch to reach state-of-the-art performance on event-based object detection - achieving an mAP of 47.2% on the Gen1 automotive dataset. At the same time, RVTs offer fast inference (<12 ms on a T4 GPU) and favorable parameter efficiency (5 times fewer than prior art). Our study brings new insights into effective design choices that can be fruitful for research beyond event-based vision.
****
Ham2Pose: Animating Sign Language Notation Into Pose Sequences
Rotem Shalev Arkushin, Amit Moryossef, Ohad Fried; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21046-21056
Translating spoken languages into Sign languages is necessary for open communication between the hearing and hearing-impaired communities. To achieve this goal, we propose the first method for animating a text written in HamNoSys, a lexical Sign language notation, into signed pose sequences. As HamNoSys is universal by design, our proposed method offers a generic solution invariant to the target Sign language. Our method gradually generates pose predictions using transformer encoders that create meaningful representations of the text and poses while considering their spatial and temporal information. We use weak supervision for the training process and show that our method succeeds in learning from partial and inaccurate data. Additionally, we offer a new distance measurement that considers missing keypoints, to measure the distance between pose sequences using DTW-MJE. We validate its correctness using AUTSL, a large-scale Sign language dataset, show that it measures the distance between pose sequences more accurately than existing measurements, and use it to assess the quality of our generated pose sequences. Code for the data pre-processing, the model, and the distance measurement is publicly released for future research.
****
Open-Set Likelihood Maximization for Few-Shot Learning
Malik Boudiaf, Etienne Bennequin, Myriam Tami, Antoine Toubhans, Pablo Piantanida, Celine Hudelot, Ismail Ben Ayed; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24007-24016
We tackle the Few-Shot Open-Set Recognition (FSOSR) problem, i.e. classifying instances among a set of classes for which we only have a few labeled samples, while simultaneously detecting instances that do not belong to any known class. We explore the popular transductive setting, which leverages the unlabelled query instances at inference. Motivated by the observation that existing transductive methods perform poorly in open-set scenarios, we propose a generalization of the maximum likelihood principle, in which latent scores down-weighing the influence of potential outliers are introduced alongside the usual parametric model. Our formulation embeds supervision constraints from the support set and additional penalties discouraging overconfident predictions on the query set. We proceed with a block-coordinate descent, with the latent scores and parametric model co-optimized alternately, thereby benefiting from each other. We call our resulting formulation Open-Set Likelihood Optimization (OSLO). OSLO is interpretable and fully modular; it can be applied on top of any pre-trained model seamlessly. Through extensive experiments, we show that our method surpasses existing inductive and transductive methods on both aspects of open-set recognition, namely inlier classification and outlier detection. Code is available at https://github.com/ebennequin/few-shot-open-set.
****
DiGeo: Discriminative Geometry-Aware Learning for Generalized Few-Shot Object Detection
Jiawei Ma, Yulei Niu, Jincheng Xu, Shiyuan Huang, Guangxing Han, Shih-Fu Chang;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3208-3218
Generalized few-shot object detection aims to achieve precise detection on both base classes with abundant annotations and novel classes with limited training data. Existing approaches enhance few-shot generalization with the sacrifice of base-class performance, or maintain high precision in base-class detection with limited improvement in novel-class adaptation. In this paper, we point out the reason is insufficient Discriminative feature learning for all of the classes. As such, we propose a new training framework, DiGeo, to learn Geometry-aware features of inter-class separation and intra-class compactness. To guide the separation of feature clusters, we derive an offline simplex equiangular tight frame (ETF) classifier whose weights serve as class centers and are maximally and equally separated. To tighten the cluster for each class, we include adaptive class-specific margins into the classification loss and encourage the features close to the class centers. Experimental studies on two few-shot benchmark datasets (PASCAL VOC, MSCOCO) and one long-tail dataset (LVIS) demonstrate that, with a single model, our method can effectively improve generalization on novel classes without hurting the detection of base classes.
****
Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation
Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, Wei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24668-24677
Distilled student models in teacher-student architectures are widely considered for computational-effective deployment in real-time applications and edge devices. However, there is a higher risk of student models to encounter adversarial attacks at the edge. Popular enhancing schemes such as adversarial training have limited performance on compressed networks. Thus, recent studies concern about adversarial distillation (AD) that aims to inherit not only prediction accuracy but also adversarial robustness of a robust teacher model under the paradigm of robust optimization. In the min-max framework of AD, existing AD methods generally use fixed supervision information from the teacher model to guide the inner optimization for knowledge distillation which often leads to an overcorrection towards model smoothness. In this paper, we propose an adaptive adversarial distillation (AdaAD) that involves the teacher model in the knowledge optimization process in a way interacting with the student model to adaptively search for the inner results. Comparing with state-of-the-art methods, the proposed AdaAD can significantly boost both the prediction accuracy and adversarial robustness of student models in most scenarios. In particular, the ResNet-18 model trained by AdaAD achieves top-rank performance (54.23% robust accuracy) on RobustBench under AutoAttack.
****
METransformer: Radiology Report Generation by Transformer With Multiple Learnable Expert Tokens
Zhanyu Wang, Lingqiao Liu, Lei Wang, Luping Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11558-11567
In clinical scenarios, multi-specialist consultation could significantly benefit the diagnosis, especially for intricate cases. This inspires us to explore a "multi-expert joint diagnosis" mechanism to upgrade the existing "single expert" framework commonly seen in the current literature. To this end, we propose METransformer, a method to realize this idea with a transformer-based backbone. The key design of our method is the introduction of multiple learnable "expert" tokens into both the transformer encoder and decoder. In the encoder, each expert token interacts with both vision tokens and other expert tokens to learn to attend different image regions for image representation. These expert tokens are encouraged to capture complementary information by an orthogonal loss that minimizes their overlap. In the decoder, each attended expert token guides the cross-attention between input words and visual tokens, thus influencing the generated report.

A metrics-based expert voting strategy is further developed to generate the final report. By the multi-experts concept, our model enjoys the merits of an ensemble-based approach but through a manner that is computationally more efficient and supports more sophisticated interactions among experts. Experimental results demonstrate the promising performance of our proposed model on two widely used benchmarks. Last but not least, the framework-level innovation makes our work ready to incorporate advances on existing "single-expert" models to further improve its performance.
****
PixHt-Lab: Pixel Height Based Light Effect Generation for Image Compositing
Yichen Sheng, Jianming Zhang, Julien Philip, Yannick Hold-Geoffroy, Xin Sun, He Zhang, Lu Ling, Bedrich Benes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16643-16653
Lighting effects such as shadows or reflections are key in making synthetic images realistic and visually appealing. To generate such effects, traditional computer graphics uses a physically-based renderer along with 3D geometry. To compensate for the lack of geometry in 2D Image compositing, recent deep learning-based approaches introduced a pixel height representation to generate soft shadows and reflections. However, the lack of geometry limits the quality of the generated soft shadows and constrains reflections to pure specular ones. We introduce PixHt-Lab, a system leveraging an explicit mapping from pixel height representation to 3D space. Using this mapping, PixHt-Lab reconstructs both the cutout and background geometry and renders realistic, diverse, lighting effects for image compositing. Given a surface with physically-based materials, we can render reflections with varying glossiness. To generate more realistic soft shadows, we further propose to use 3D-aware buffer channels to guide a neural renderer. Both quantitative and qualitative evaluations demonstrate that PixHt-Lab significantly improves soft shadow generation.
****
A Soma Segmentation Benchmark in Full Adult Fly Brain
Xiaoyu Liu, Bo Hu, Mingxing Li, Wei Huang, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7402-7411
Neuron reconstruction in a full adult fly brain from high-resolution electron microscopy (EM) data is regarded as a cornerstone for neuroscientists to explore how neurons inspire intelligence. As the central part of neurons, somas in the full brain indicate the origin of neurogenesis and neural functions. However, due to the absence of EM datasets specifically annotated for somas, existing deep learning-based neuron reconstruction methods cannot directly provide accurate soma distribution and morphology. Moreover, full brain neuron reconstruction remains extremely time-consuming due to the unprecedentedly large size of EM data. In this paper, we develop an efficient soma reconstruction method for obtaining accurate soma distribution and morphology information in a full adult fly brain. To this end, we first make a high-resolution EM dataset with fine-grained 3D manual annotations on somas. Relying on this dataset, we propose an efficient, two-stage deep learning algorithm for predicting accurate locations and boundaries of 3D soma instances. Further, we deploy a parallelized, high-throughput data processing pipeline for executing the above algorithm on the full brain. Finally, we provide quantitative and qualitative benchmark comparisons on the testset to validate the superiority of the proposed method, as well as preliminary statistics of the reconstructed somas in the full adult fly brain from the biological perspective. We release our code and dataset at https://github.com/liuxy1103/EMADS.
****
RGB No More: Minimally-Decoded JPEG Vision Transformers
Jeongsoo Park, Justin Johnson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22334-22346
Most neural networks for computer vision are designed to infer using RGB images. However, these RGB images are commonly encoded in JPEG before saving to disk; decoding them imposes an unavoidable overhead for RGB networks. Instead, our work focuses on training Vision Transformers (ViT) directly from the encoded feature

s of JPEG. This way, we can avoid most of the decoding overhead, accelerating da
ta load. Existing works have studied this aspect but they focus on CNNs. Due to
how these encoded features are structured, CNNs require heavy modification to th
eir architecture to accept such data. Here, we show that this is not the case fo
r ViTs. In addition, we tackle data augmentation directly on these encoded featu
res, which to our knowledge, has not been explored in-depth for training in this
 setting. With these two improvements -- ViT and data augmentation -- we show th
at our ViT-Ti model achieves up to 39.2% faster training and 17.9% faster infere
nce with no accuracy loss compared to the RGB counterpart.
****
Revealing the Dark Secrets of Masked Image Modeling
Zhenda Xie, Zigang Geng, Jingcheng Hu, Zheng Zhang, Han Hu, Yue Cao; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 14475-14485
Masked image modeling (MIM) as pre-training is shown to be effective for numerou
s vision downstream tasks, but how and where MIM works remain unclear. In this p
aper, we compare MIM with the long-dominant supervised pre-trained models from t
wo perspectives, the visualizations and the experiments, to uncover their key re
presentational differences. From the visualizations, we find that MIM brings loc
ality inductive bias to all layers of the trained models, but supervised models
tend to focus locally at lower layers but more globally at higher layers. That m
ay be the reason why MIM helps Vision Transformers that have a very large recept
ive field to optimize. Using MIM, the model can maintain a large diversity on at
tention heads in all layers. But for supervised models, the diversity on attenti
on heads almost disappears from the last three layers and less diversity harms t
he fine-tuning performance. From the experiments, we find that MIM models can pe
rform significantly better on geometric and motion tasks with weak semantics or
fine-grained classification tasks, than their supervised counterparts. Without b
ells and whistles, a standard MIM pre-trained SwinV2-L could achieve state-of-th
e-art performance on pose estimation (78.9 AP on COCO test-dev and 78.0 AP on Cr
owdPose), depth estimation (0.287 RMSE on NYUv2 and 1.966 RMSE on KITTI), and vi
deo object tracking (70.7 SUC on LaSOT). For the semantic understanding datasets
 where the categories are sufficiently covered by the supervised pre-training, M
IM models can still achieve highly competitive transfer performance. With a deep
er understanding of MIM, we hope that our work can inspire new and solid researc
h in this direction. Code will be available at https://github.com/zdaxie/MIM-Dar
kSecrets.
****
Fine-Grained Classification With Noisy Labels
Qi Wei, Lei Feng, Haoliang Sun, Ren Wang, Chenhui Guo, Yilong Yin; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
3, pp. 11651-11660
Learning with noisy labels (LNL) aims to ensure model generalization given a lab
el-corrupted training set. In this work, we investigate a rarely studied scenari
o of LNL on fine-grained datasets (LNL-FG), which is more practical and challeng
ing as large inter-class ambiguities among fine-grained classes cause more noisy
 labels. We empirically show that existing methods that work well for LNL fail t
o achieve satisfying performance for LNL-FG, arising the practical need of effec
tive solutions for LNL-FG. To this end, we propose a novel framework called stoc
hastic noise-tolerated supervised contrastive learning (SNSCL) that confronts la
bel noise by encouraging distinguishable representation. Specifically, we design
 a noise-tolerated supervised contrastive learning loss that incorporates a weig
ht-aware mechanism for noisy label correction and selectively updating momentum
queue lists. By this mechanism, we mitigate the effects of noisy anchors and avo
id inserting noisy labels into the momentum-updated queue. Besides, to avoid man
ually-defined augmentation strategies in contrastive learning, we propose an eff
icient stochastic module that samples feature embeddings from a generated distri
bution, which can also enhance the representation ability of deep models. SNSCL
is general and compatible with prevailing robust LNL strategies to improve their
 performance for LNL-FG. Extensive experiments demonstrate the effectiveness of

SNSCL.
****

CaPriDe Learning: Confidential and Private Decentralized Learning Based on Encryption-Friendly Distillation Loss
Nurbek Tastan, Karthik Nandakumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8084-8092
Large volumes of data required to train accurate deep neural networks (DNNs) are seldom available with any single entity. Often, privacy concerns and stringent data regulations prevent entities from sharing data with each other or with a third-party learning service provider. While cross-silo federated learning (FL) allows collaborative learning of large DNNs without sharing the data itself, most existing cross-silo FL algorithms have an unacceptable utility-privacy trade-off. In this work, we propose a framework called Confidential and Private Decentralized (CaPriDe) learning, which optimally leverages the power of fully homomorphic encryption (FHE) to enable collaborative learning without compromising on the confidentiality and privacy of data. In CaPriDe learning, participating entities release their private data in an encrypted form allowing other participants to perform inference in the encrypted domain. The crux of CaPriDe learning is mutual knowledge distillation between multiple local models through a novel distillation loss, which is an approximation of the Kullback-Leibler (KL) divergence between the local predictions and encrypted inferences of other participants on the same data that can be computed in the encrypted domain. Extensive experiments on three datasets show that CaPriDe learning can improve the accuracy of local models without any central coordination, provide strong guarantees of data confidentiality and privacy, and has the ability to handle statistical heterogeneity. Constraints on the model architecture (arising from the need to be FHE-friendly), limited scalability, and computational complexity of encrypted domain inference are the main limitations of the proposed approach. The code can be found at https://github.com/tnurbek/capride-learning.
****

Hybrid Active Learning via Deep Clustering for Video Action Detection
Aayush J. Rana, Yogesh S. Rawat; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18867-18877
In this work, we focus on reducing the annotation cost for video action detection which requires costly frame-wise dense annotations. We study a novel hybrid active learning (AL) strategy which performs efficient labeling using both intra-sample and inter-sample selection. The intra-sample selection leads to labeling of fewer frames in a video as opposed to inter-sample selection which operates at video level. This hybrid strategy reduces the annotation cost from two different aspects leading to significant labeling cost reduction. The proposed approach utilize Clustering-Aware Uncertainty Scoring (CLAUS), a novel label acquisition strategy which relies on both informativeness and diversity for sample selection. We also propose a novel Spatio-Temporal Weighted (STeW) loss formulation, which helps in model training under limited annotations. The proposed approach is evaluated on UCF-101-24 and J-HMDB-21 datasets demonstrating its effectiveness in significantly reducing the annotation cost where it consistently outperforms other baselines. Project details available at https://sites.google.com/view/activesparselabeling/home
****

Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network
Zhengxin Pan, Fangyu Wu, Bailing Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19275-19284
Current state-of-the-art image-text matching methods implicitly align the visual-semantic fragments, like regions in images and words in sentences, and adopt cross-attention mechanism to discover fine-grained cross-modal semantic correspondence. However, the cross-attention mechanism may bring redundant or irrelevant region-word alignments, degenerating retrieval accuracy and limiting efficiency. Although many researchers have made progress in mining meaningful alignments and thus improving accuracy, the problem of poor efficiency remains unresolved. In this work, we propose to learn fine-grained image-text matching from the perspec

tive of information coding. Specifically, we suggest a coding framework to expla
in the fragments aligning process, which provides a novel view to reexamine the
cross-attention mechanism and analyze the problem of redundant alignments. Based
 on this framework, a Cross-modal Hard Aligning Network (CHAN) is designed, whic
h comprehensively exploits the most relevant region-word pairs and eliminates al
l other alignments. Extensive experiments conducted on two public datasets, MS-C
OCO and Flickr30K, verify that the relevance of the most associated word-region
pairs is discriminative enough as an indicator of the image-text similarity, wit
h superior accuracy and efficiency over the state-of-the-art approaches on the b
idirectional image and text retrieval tasks. Our code will be available at https
://github.com/ppanzx/CHAN.
****
Sparsifiner: Learning Sparse Instance-Dependent Attention for Efficient Vision T
ransformers
Cong Wei, Brendan Duke, Ruowei Jiang, Parham Aarabi, Graham W. Taylor, Florian S
hkurti; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2023, pp. 22680-22689
Vision Transformers (ViT) have shown competitive advantages in terms of performa
nce compared to convolutional neural networks (CNNs), though they often come wit
h high computational costs. To this end, previous methods explore different atte
ntion patterns by limiting a fixed number of spatially nearby tokens to accelera
te the ViT's multi-head self-attention (MHSA) operations. However, such structur
ed attention patterns limit the token-to-token connections to their spatial rele
vance, which disregards learned semantic connections from a full attention mask.
 In this work, we propose an approach to learn instance-dependent attention patt
erns, by devising a lightweight connectivity predictor module that estimates the
 connectivity score of each pair of tokens. Intuitively, two tokens have high co
nnectivity scores if the features are considered relevant either spatially or se
mantically. As each token only attends to a small number of other tokens, the bi
narized connectivity masks are often very sparse by nature and therefore provide
 the opportunity to reduce network FLOPs via sparse computations. Equipped with
the learned unstructured attention pattern, sparse attention ViT (Sparsifiner) p
roduces a superior Pareto frontier between FLOPs and top-1 accuracy on ImageNet
compared to token sparsity. Our method reduces 48%   69% FLOPs of MHSA while the
 accuracy drop is within 0.4%. We also show that combining attention and token s
parsity reduces ViT FLOPs by over 60%.
****
Structured Sparsity Learning for Efficient Video Super-Resolution
Bin Xia, Jingwen He, Yulun Zhang, Yitong Wang, Yapeng Tian, Wenming Yang, Luc Va
n Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re
cognition (CVPR), 2023, pp. 22638-22647
The high computational costs of video super-resolution (VSR) models hinder their
 deployment on resource-limited devices, e.g., smartphones and drones. Existing
VSR models contain considerable redundant filters, which drag down the inference
 efficiency. To prune these unimportant filters, we develop a structured pruning
 scheme called Structured Sparsity Learning (SSL) according to the properties of
 VSR. In SSL, we design pruning schemes for several key components in VSR models
, including residual blocks, recurrent networks, and upsampling networks. Specif
ically, we develop a Residual Sparsity Connection (RSC) scheme for residual bloc
ks of recurrent networks to liberate pruning restrictions and preserve the resto
ration information. For upsampling networks, we design a pixel-shuffle pruning s
cheme to guarantee the accuracy of feature channel-space conversion. In addition
, we observe that pruning error would be amplified as the hidden states propagat
e along with recurrent networks. To alleviate the issue, we design Temporal Fine
tuning (TF). Extensive experiments show that SSL can significantly outperform re
cent methods quantitatively and qualitatively. The code is available at https://
github.com/Zj-BinXia/SSL.
****
CAP: Robust Point Cloud Classification via Semantic and Structural Modeling
Daizong Ding, Erling Jiang, Yuanmin Huang, Mi Zhang, Wenxuan Li, Min Yang; Proce

edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12260-12270
Recently, deep neural networks have shown great success on 3D point cloud classification tasks, which simultaneously raises the concern of adversarial attacks that cause severe damage to real-world applications. Moreover, defending against adversarial examples in point cloud data is extremely difficult due to the emergence of various attack strategies. In this work, with the insight of the fact that the adversarial examples in this task still preserve the same semantic and structural information as the original input, we design a novel defense framework for improving the robustness of existing classification models, which consists of two main modules: the attention-based pooling and the dynamic contrastive learning. In addition, we also develop an algorithm to theoretically certify the robustness of the proposed framework. Extensive empirical results on two datasets and three classification models show the robustness of our approach against various attacks, e.g., the averaged attack success rate of PointNet decreases from 70.2% to 2.7% on the ModelNet40 dataset under 9 common attacks.
****
"Seeing" Electric Network Frequency From Events
Lexuan Xu, Guang Hua, Haijian Zhang, Lei Yu, Ning Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18022-18031
Most of the artificial lights fluctuate in response to the grid's alternating current and exhibit subtle variations in terms of both intensity and spectrum, providing the potential to estimate the Electric Network Frequency (ENF) from conventional frame-based videos. Nevertheless, the performance of Video-based ENF (V-ENF) estimation largely relies on the imaging quality and thus may suffer from significant interference caused by non-ideal sampling, motion, and extreme lighting conditions. In this paper, we show that the ENF can be extracted without the above limitations from a new modality provided by the so-called event camera, a neuromorphic sensor that encodes the light intensity variations and asynchronously emits events with extremely high temporal resolution and high dynamic range. Specifically, we first formulate and validate the physical mechanism for the ENF captured in events, and then propose a simple yet robust Event-based ENF (E-ENF) estimation method through mode filtering and harmonic enhancement. Furthermore, we build an Event-Video ENF Dataset (EV-ENFD) that records both events and videos in diverse scenes. Extensive experiments on EV-ENFD demonstrate that our proposed E-ENF method can extract more accurate ENF traces, outperforming the conventional V-ENF by a large margin, especially in challenging environments with object motions and extreme lighting conditions. The code and dataset are available at https://github.com/xlx-creater/E-ENF.
****
MMVC: Learned Multi-Mode Video Compression With Block-Based Prediction Mode Selection and Density-Adaptive Entropy Coding
Bowen Liu, Yu Chen, Rakesh Chowdary Machineni, Shiyu Liu, Hun-Seok Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18487-18496
Learning-based video compression has been extensively studied over the past years, but it still has limitations in adapting to various motion patterns and entropy models. In this paper, we propose multi-mode video compression (MMVC), a block wise mode ensemble deep video compression framework that selects the optimal mode for feature domain prediction adapting to different motion patterns. Proposed multi-modes include ConvLSTM-based feature domain prediction, optical flow conditioned feature domain prediction, and feature propagation to address a wide range of cases from static scenes without apparent motions to dynamic scenes with a moving camera. We partition the feature space into blocks for temporal prediction in spatial block-based representations. For entropy coding, we consider both dense and sparse post-quantization residual blocks, and apply optional run-length coding to sparse residuals to improve the compression rate. In this sense, our method uses a dual-mode entropy coding scheme guided by a binary density map, which offers significant rate reduction surpassing the extra cost of transmittin

g the binary selection map. We validate our scheme with some of the most popular benchmarking datasets. Compared with state-of-the-art video compression schemes and standard codecs, our method yields better or competitive results measured with PSNR and MS-SSIM.
****
Visual-Tactile Sensing for In-Hand Object Reconstruction
Wenqiang Xu, Zhenjun Yu, Han Xue, Ruolin Ye, Siqiong Yao, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8803-8812
Tactile sensing is one of the modalities human rely on heavily to perceive the world. Working with vision, this modality refines local geometry structure, measures deformation at contact area, and indicates hand-object contact state. With the availability of open-source tactile sensors such as DIGIT, research on visual-tactile learning is becoming more accessible and reproducible. Leveraging this tactile sensor, we propose a novel visual-tactile in-hand object reconstruction framework VTacO, and extend it to VTacOH for hand-object reconstruction. Since our method can support both rigid and deformable object reconstruction, and no existing benchmark are proper for the goal. We propose a simulation environment, VT-Sim, which supports to generate hand-object interaction for both rigid and deformable objects. With VT-Sim, we generate a large-scale training dataset, and evaluate our method on it. Extensive experiments demonstrate that our proposed method can outperform the previous baseline methods qualitatively and quantitatively. Finally, we directly apply our model trained in simulation to various real-world test cases, which display qualitative results. Codes, models, simulation environment, datasets will be publicly available.
****
vMAP: Vectorised Object Mapping for Neural Field SLAM
Xin Kong, Shikun Liu, Marwan Taher, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 952-961
We present vMAP, an object-level dense SLAM system using neural field representations. Each object is represented by a small MLP, enabling efficient, watertight object modelling without the need for 3D priors. As an RGB-D camera browses a scene with no prior information, vMAP detects object instances on-the-fly, and dynamically adds them to its map. Specifically, thanks to the power of vectorised training, vMAP can optimise as many as 50 individual objects in a single scene, with an extremely efficient training speed of 5Hz map update. We experimentally demonstrate significantly improved scene-level and object-level reconstruction quality compared to prior neural field SLAM systems. Project page: https://kxhit.github.io/vMAP.
****
Images Speak in Images: A Generalist Painter for In-Context Visual Learning
Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, Tiejun Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6830-6839
In-context learning, as a new paradigm in NLP, allows the model to rapidly adapt to various tasks with only a handful of prompts and examples. But in computer vision, the difficulties for in-context learning lie in that tasks vary significantly in the output representations, thus it is unclear how to define the general-purpose task prompts that the vision model can understand and transfer to out-of-domain tasks. In this work, we present Painter, a generalist model which addresses these obstacles with an "image"-centric solution, that is, to redefine the output of core vision tasks as images, and specify task prompts as also images. With this idea, our training process is extremely simple, which performs standard masked image modeling on the stitch of input and output image pairs. This makes the model capable of performing tasks conditioned on visible image patches. Thus, during inference, we can adopt a pair of input and output images from the same task as the input condition, to indicate which task to perform. Without bells and whistles, our generalist Painter can achieve competitive performance compared to well-established task-specific models, on seven representative vision task

s ranging from high-level visual understanding to low-level image processing. In addition, Painter significantly outperforms recent generalist models on several challenging tasks.
****
Omni Aggregation Networks for Lightweight Image Super-Resolution
Hang Wang, Xuanhong Chen, Bingbing Ni, Yutian Liu, Jinfan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22378-22387
While lightweight ViT framework has made tremendous progress in image super-resolution, its uni-dimensional self-attention modeling, as well as homogeneous aggregation scheme, limit its effective receptive field (ERF) to include more comprehensive interactions from both spatial and channel dimensions. To tackle these drawbacks, this work proposes two enhanced components under a new Omni-SR architecture. First, an Omni Self-Attention (OSA) paradigm is proposed based on dense interaction principle, which can simultaneously model pixel-interaction from both spatial and channel dimensions, mining the potential correlations across omni-axis (i.e., spatial and channel). Coupling with mainstream window partitioning strategies, OSA can achieve superior performance with compelling computational budgets. Second, a multi-scale interaction scheme is proposed to mitigate sub-optimal ERF (i.e., premature saturation) in shallow models, which facilitates local propagation and meso-/global-scale interactions, rendering a omni-scale aggregation building block. Extensive experiments demonstrate that Omni-SR achieves record-high performance on lightweight super-resolution benchmarks (e.g., 26.95dB@Urban100 x4 with only 792K parameters). Our code is available at https://github.com/Francis0625/Omni-SR.
****
StyLess: Boosting the Transferability of Adversarial Examples
Kaisheng Liang, Bin Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8163-8172
Adversarial attacks can mislead deep neural networks (DNNs) by adding imperceptible perturbations to benign examples. The attack transferability enables adversarial examples to attack black-box DNNs with unknown architectures or parameters, which poses threats to many real-world applications. We find that existing transferable attacks do not distinguish between style and content features during optimization, limiting their attack transferability. To improve attack transferability, we propose a novel attack method called style-less perturbation (StyLess). Specifically, instead of using a vanilla network as the surrogate model, we advocate using stylized networks, which encode different style features by perturbing an adaptive instance normalization. Our method can prevent adversarial examples from using non-robust style features and help generate transferable perturbations. Comprehensive experiments show that our method can significantly improve the transferability of adversarial examples. Furthermore, our approach is generic and can outperform state-of-the-art transferable attacks when combined with other attack techniques.
****
Local-to-Global Registration for Bundle-Adjusting Neural Radiance Fields
Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, Fei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8264-8273
Neural Radiance Fields (NeRF) have achieved photorealistic novel views synthesis; however, the requirement of accurate camera poses limits its application. Despite analysis-by-synthesis extensions for jointly learning neural 3D representations and registering camera frames exist, they are susceptible to suboptimal solutions if poorly initialized. We propose L2G-NeRF, a Local-to-Global registration method for bundle-adjusting Neural Radiance Fields: first, a pixel-wise flexible alignment, followed by a frame-wise constrained parametric alignment. Pixel-wise local alignment is learned in an unsupervised way via a deep network which optimizes photometric reconstruction errors. Frame-wise global alignment is performed using differentiable parameter estimation solvers on the pixel-wise correspondences to find a global transformation. Experiments on synthetic and real-world

data show that our method outperforms the current state-of-the-art in terms of high-fidelity reconstruction and resolving large camera pose misalignment. Our module is an easy-to-use plugin that can be applied to NeRF variants and other neural field applications.
****
Uncertainty-Aware Optimal Transport for Semantically Coherent Out-of-Distribution Detection
Fan Lu, Kai Zhu, Wei Zhai, Kecheng Zheng, Yang Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3282-3291
Semantically coherent out-of-distribution (SCOOD) detection aims to discern outliers from the intended data distribution with access to unlabeled extra set. The coexistence of in-distribution and out-of-distribution samples will exacerbate the model overfitting when no distinction is made. To address this problem, we propose a novel uncertainty-aware optimal transport scheme. Our scheme consists of an energy-based transport (ET) mechanism that estimates the fluctuating cost of uncertainty to promote the assignment of semantic-agnostic representation, and an inter-cluster extension strategy that enhances the discrimination of semantic property among different clusters by widening the corresponding margin distance. Furthermore, a T-energy score is presented to mitigate the magnitude gap between the parallel transport and classifier branches. Extensive experiments on two standard SCOOD benchmarks demonstrate the above-par OOD detection performance, outperforming the state-of-the-art methods by a margin of 27.69% and 34.4% on FPR@95, respectively.
****
FJMP: Factorized Joint Multi-Agent Motion Prediction Over Learned Directed Acyclic Interaction Graphs
Luke Rowe, Martin Ethier, Eli-Henry Dykhne, Krzysztof Czarnecki; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13745-13755
Predicting the future motion of road agents is a critical task in an autonomous driving pipeline. In this work, we address the problem of generating a set of scene-level, or joint, future trajectory predictions in multi-agent driving scenarios. To this end, we propose FJMP, a Factorized Joint Motion Prediction framework for multi-agent interactive driving scenarios. FJMP models the future scene interaction dynamics as a sparse directed interaction graph, where edges denote explicit interactions between agents. We then prune the graph into a directed acyclic graph (DAG) and decompose the joint prediction task into a sequence of marginal and conditional predictions according to the partial ordering of the DAG, where joint future trajectories are decoded using a directed acyclic graph neural network (DAGNN). We conduct experiments on the INTERACTION and Argoverse 2 datasets and demonstrate that FJMP produces more accurate and scene-consistent joint trajectory predictions than non-factorized approaches, especially on the most interactive and kinematically interesting agents. FJMP ranks 1st on the multi-agent test leaderboard of the INTERACTION dataset.
****
Exploring the Effect of Primitives for Compositional Generalization in Vision-and-Language
Chuanhao Li, Zhen Li, Chenchen Jing, Yunde Jia, Yuwei Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19092-19101
Compositionality is one of the fundamental properties of human cognition (Fodor & Pylyshyn, 1988). Compositional generalization is critical to simulate the compositional capability of humans, and has received much attention in the vision-and-language (V&L) community. It is essential to understand the effect of the primitives, including words, image regions, and video frames, to improve the compositional generalization capability. In this paper, we explore the effect of primitives for compositional generalization in V&L. Specifically, we present a self-supervised learning based framework that equips V&L methods with two characteristics: semantic equivariance and semantic invariance. With the two characteristics,

the methods understand primitives by perceiving the effect of primitive changes on sample semantics and ground-truth. Experimental results on two tasks: tempor al video grounding and visual question answering, demonstrate the effectiveness of our framework.
****
Correlational Image Modeling for Self-Supervised Visual Pre-Training
Wei Li, Jiahao Xie, Chen Change Loy; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2023, pp. 15105-15115
We introduce Correlational Image Modeling (CIM), a novel but surprisingly effect ive approach to self-supervised visual pre-training. Our CIM performs a simple p retext task: we randomly crop image regions (exemplar) from an input image (cont ext) and predict correlation maps between the exemplars and the context. Three k ey designs enable correlational image modeling as a nontrivial and meaningful se lf-supervisory task. First, to generate useful exemplar-context pairs, we consid er cropping image regions with various scales, shapes, rotations, and transforma tions. Second, we employ a bootstrap learning framework that involves online and target networks. During pre-training, the former takes exemplars as inputs whil e the latter converts the context. Third, we model the output correlation maps v ia a simple cross-attention block, within which the context serves as queries an d the exemplars offer values and keys. We show that CIM performs on par or bette r than the current state of the art on self-supervised and transfer benchmarks.
****
DC2: Dual-Camera Defocus Control by Learning To Refocus
Hadi Alzayer, Abdullah Abuolaim, Leung Chun Chan, Yang Yang, Ying Chen Lou, Jia-Bin Huang, Abhishek Kar; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2023, pp. 21488-21497
Smartphone cameras today are increasingly approaching the versatility and qualit y of professional cameras through a combination of hardware and software advance ments. However, fixed aperture remains a key limitation, preventing users from c ontrolling the depth of field (DoF) of captured images. At the same time, many s martphones now have multiple cameras with different fixed apertures - specifical ly, an ultra-wide camera with wider field of view and deeper DoF and a higher re solution primary camera with shallower DoF. In this work, we propose DC^2, a sys tem for defocus control for synthetically varying camera aperture, focus distanc e and arbitrary defocus effects by fusing information from such a dual-camera sy stem. Our key insight is to leverage real-world smartphone camera dataset by usi ng image refocus as a proxy task for learning to control defocus. Quantitative a nd qualitative evaluations on real-world data demonstrate our system's efficacy where we outperform state-of-the-art on defocus deblurring, bokeh rendering, and image refocus. Finally, we demonstrate creative post-capture defocus control en abled by our method, including tilt-shift and content-based defocus effects.
****
MISC210K: A Large-Scale Dataset for Multi-Instance Semantic Correspondence
Yixuan Sun, Yiwen Huang, Haijing Guo, Yuzhou Zhao, Runmin Wu, Yizhou Yu, Weifeng Ge, Wenqiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2023, pp. 7121-7130
Semantic correspondence have built up a new way for object recognition. However current single-object matching schema can be hard for discovering commonalities for a category and far from the real-world recognition tasks. To fill this gap, we design the multi-instance semantic correspondence task which aims at construc ting the correspondence between multiple objects in an image pair. To support th is task, we build a multi-instance semantic correspondence (MISC) dataset from C OCO Detection 2017 task called MISC210K. We construct our dataset as three steps : (1) category selection and data cleaning; (2) keypoint design based on 3D mode ls and object description rules; (3) human-machine collaborative annotation. Fol lowing these steps, we select 34 classes of objects with 4,812 challenging image s annotated via a well designed semi-automatic workflow, and finally acquire 218 ,179 image pairs with instance masks and instance-level keypoint pairs annotated . We design a dual-path collaborative learning pipeline to train instance-level co-segmentation task and fine-grained level correspondence task together. Benchm

ark evaluation and further ablation results with detailed analysis are provided with three future directions proposed. Our project is available on https://github.com/YXSUNMADMAX/MISC210K.
****
Self-Supervised Implicit Glyph Attention for Text Recognition
Tongkun Guan, Chaochen Gu, Jingzheng Tu, Xue Yang, Qi Feng, Yudi Zhao, Wei Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15285-15294
The attention mechanism has become the de facto module in scene text recognition (STR) methods, due to its capability of extracting character-level representations. These methods can be summarized into implicit attention based and supervised attention based, depended on how the attention is computed, i.e., implicit attention and supervised attention are learned from sequence-level text annotations and character-level bounding box annotations, respectively. Implicit attention, as it may extract coarse or even incorrect spatial regions as character attention, is prone to suffering from an alignment-drifted issue. Supervised attention can alleviate the above issue, but it is category-specific, which requires extra laborious character-level bounding box annotations and would be memory-intensive when the number of character categories is large. To address the aforementioned issues, we propose a novel attention mechanism for STR, self-supervised implicit glyph attention (SIGA). SIGA delineates the glyph structures of text images by jointly self-supervised text segmentation and implicit attention alignment, which serve as the supervision to improve attention correctness without extra character-level annotations. Experimental results demonstrate that SIGA performs consistently and significantly better than previous attention-based STR methods, in terms of both attention correctness and final recognition performance on publicly available context benchmarks and our contributed contextless benchmarks.
****
ACL-SPC: Adaptive Closed-Loop System for Self-Supervised Point Cloud Completion
Sangmin Hong, Mohsen Yavartanoo, Reyhaneh Neshatavar, Kyoung Mu Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9435-9444
Point cloud completion addresses filling in the missing parts of a partial point cloud obtained from depth sensors and generating a complete point cloud. Although there has been steep progress in the supervised methods on the synthetic point cloud completion task, it is hardly applicable in real-world scenarios due to the domain gap between the synthetic and real-world datasets or the requirement of prior information. To overcome these limitations, we propose a novel self-supervised framework ACL-SPC for point cloud completion to train and test on the same data. ACL-SPC takes a single partial input and attempts to output the complete point cloud using an adaptive closed-loop (ACL) system that enforces the output same for the variation of an input. We evaluate our ACL-SPC on various datasets to prove that it can successfully learn to complete a partial point cloud as the first self-supervised scheme. Results show that our method is comparable with unsupervised methods and achieves superior performance on the real-world dataset compared to the supervised methods trained on the synthetic dataset. Extensive experiments justify the necessity of self-supervised learning and the effectiveness of our proposed method for the real-world point cloud completion task. The code is publicly available from this link.
****
MAGE: MAsked Generative Encoder To Unify Representation Learning and Image Synthesis
Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, Dilip Krishnan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2142-2152
Generative modeling and representation learning are two key tasks in computer vision. However, these models are typically trained independently, which ignores the potential for each task to help the other, and leads to training and model maintenance overheads. In this work, we propose MAsked Generative Encoder (MAGE), the first framework to unify SOTA image generation and self-supervised represent

ation learning. Our key insight is that using variable masking ratios in masked image modeling pre-training can allow generative training (very high masking ratio) and representation learning (lower masking ratio) under the same training framework. Inspired by previous generative models, MAGE uses semantic tokens learned by a vector-quantized GAN at inputs and outputs, combining this with masking. We can further improve the representation by adding a contrastive loss to the encoder output. We extensively evaluate the generation and representation learning capabilities of MAGE. On ImageNet-1K, a single MAGE ViT-L model obtains 9.10 FID in the task of class-unconditional image generation and 78.9% top-1 accuracy for linear probing, achieving state-of-the-art performance in both image generation and representation learning. Code is available at https://github.com/LTH14/mage.
****
Focus on Details: Online Multi-Object Tracking With Diverse Fine-Grained Representation
Hao Ren, Shoudong Han, Huilin Ding, Ziwen Zhang, Hongwei Wang, Faquan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11289-11298
Discriminative representation is essential to keep a unique identifier for each target in Multiple object tracking (MOT). Some recent MOT methods extract features of the bounding box region or the center point as identity embeddings. However, when targets are occluded, these coarse-grained global representations become unreliable. To this end, we propose exploring diverse fine-grained representation, which describes appearance comprehensively from global and local perspectives. This fine-grained representation requires high feature resolution and precise semantic information. To effectively alleviate the semantic misalignment caused by indiscriminate contextual information aggregation, Flow Alignment FPN (FAFPN) is proposed for multi-scale feature alignment aggregation. It generates semantic flow among feature maps from different resolutions to transform their pixel positions. Furthermore, we present a Multi-head Part Mask Generator (MPMG) to extract fine-grained representation based on the aligned feature maps. Multiple parallel branches of MPMG allow it to focus on different parts of targets to generate local masks without label supervision. The diverse details in target masks facilitate fine-grained representation. Eventually, benefiting from a Shuffle-Group Sampling (SGS) training strategy with positive and negative samples balanced, we achieve state-of-the-art performance on MOT17 and MOT20 test sets. Even on DanceTrack, where the appearance of targets is extremely similar, our method significantly outperforms ByteTrack by 5.0% on HOTA and 5.6% on IDF1. Extensive experiments have proved that diverse fine-grained representation makes Re-ID great again in MOT.
****
DiffPose: Toward More Reliable 3D Pose Estimation
Jia Gong, Lin Geng Foo, Zhipeng Fan, Qiuhong Ke, Hossein Rahmani, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13041-13051
Monocular 3D human pose estimation is quite challenging due to the inherent ambiguity and occlusion, which often lead to high uncertainty and indeterminacy. On the other hand, diffusion models have recently emerged as an effective tool for generating high-quality images from noise. Inspired by their capability, we explore a novel pose estimation framework (DiffPose) that formulates 3D pose estimation as a reverse diffusion process. We incorporate novel designs into our DiffPose to facilitate the diffusion process for 3D pose estimation: a pose-specific initialization of pose uncertainty distributions, a Gaussian Mixture Model-based forward diffusion process, and a context-conditioned reverse diffusion process. Our proposed DiffPose significantly outperforms existing methods on the widely used pose estimation benchmarks Human3.6M and MPI-INF-3DHP. Project page: https://gongjia0208.github.io/Diffpose/.
****
Lift3D: Synthesize 3D Training Data by Lifting 2D GAN to 3D Generative Radiance Field

Leheng Li, Qing Lian, Luozhou Wang, Ningning Ma, Ying-Cong Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 332-341

This work explores the use of 3D generative models to synthesize training data for 3D vision tasks. The key requirements of the generative models are that the generated data should be photorealistic to match the real-world scenarios, and the corresponding 3D attributes should be aligned with given sampling labels. However, we find that the recent NeRF-based 3D GANs hardly meet the above requirements due to their designed generation pipeline and the lack of explicit 3D supervision. In this work, we propose Lift3D, an inverted 2D-to-3D generation framework to achieve the data generation objectives. Lift3D has several merits compared to prior methods: (1) Unlike previous 3D GANs that the output resolution is fixed after training, Lift3D can generalize to any camera intrinsic with higher resolution and photorealistic output. (2) By lifting well-disentangled 2D GAN to 3D object NeRF, Lift3D provides explicit 3D information of generated objects, thus offering accurate 3D annotations for downstream tasks. We evaluate the effectiveness of our framework by augmenting autonomous driving datasets. Experimental results demonstrate that our data generation framework can effectively improve the performance of 3D object detectors. Code: len-li.github.io/lift3d-web

****

Hunting Sparsity: Density-Guided Contrastive Learning for Semi-Supervised Semantic Segmentation

Xiaoyang Wang, Bingfeng Zhang, Limin Yu, Jimin Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3114-3123

Recent semi-supervised semantic segmentation methods combine pseudo labeling and consistency regularization to enhance model generalization from perturbation-invariant training. In this work, we argue that adequate supervision can be extracted directly from the geometry of feature space. Inspired by density-based unsupervised clustering, we propose to leverage feature density to locate sparse regions within feature clusters defined by label and pseudo labels. The hypothesis is that lower-density features tend to be under-trained compared with those densely gathered. Therefore, we propose to apply regularization on the structure of the cluster by tackling the sparsity to increase intra-class compactness in feature space. With this goal, we present a Density-Guided Contrastive Learning (DGCL) strategy to push anchor features in sparse regions toward cluster centers approximated by high-density positive keys. The heart of our method is to estimate feature density which is defined as neighbor compactness. We design a multi-scale density estimation module to obtain the density from multiple nearest-neighbor graphs for robust density modeling. Moreover, a unified training framework is proposed to combine label-guided self-training and density-guided geometry regularization to form complementary supervision on unlabeled data. Experimental results on PASCAL VOC and Cityscapes under various semi-supervised settings demonstrate that our proposed method achieves state-of-the-art performances.

****

Learning Analytical Posterior Probability for Human Mesh Recovery

Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, Weidong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8781-8791

Despite various probabilistic methods for modeling the uncertainty and ambiguity in human mesh recovery, their overall precision is limited because existing formulations for joint rotations are either not constrained to SO(3) or difficult to learn for neural networks. To address such an issue, we derive a novel analytical formulation for learning posterior probability distributions of human joint rotations conditioned on bone directions in a Bayesian manner, and based on this, we propose a new posterior-guided framework for human mesh recovery. We demonstrate that our framework is not only superior to existing SOTA baselines on multiple benchmarks but also flexible enough to seamlessly incorporate with additional sensors due to its Bayesian nature. The code is available at https://github.com/NetEase-GameAI/ProPose.

****
Looking Through the Glass: Neural Surface Reconstruction Against High Specular Reflections
Jiaxiong Qiu, Peng-Tao Jiang, Yifan Zhu, Ze-Xin Yin, Ming-Ming Cheng, Bo Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20823-20833
Neural implicit methods have achieved high-quality 3D object surfaces under slight specular highlights. However, high specular reflections (HSR) often appear in front of target objects when we capture them through glasses. The complex ambiguity in these scenes violates the multi-view consistency, then makes it challenging for recent methods to reconstruct target objects correctly. To remedy this issue, we present a novel surface reconstruction framework, NeuS-HSR, based on implicit neural rendering. In NeuS-HSR, the object surface is parameterized as an implicit signed distance function (SDF). To reduce the interference of HSR, we propose decomposing the rendered image into two appearances: the target object and the auxiliary plane. We design a novel auxiliary plane module by combining physical assumptions and neural networks to generate the auxiliary plane appearance. Extensive experiments on synthetic and real-world datasets demonstrate that NeuS-HSR outperforms state-of-the-art approaches for accurate and robust target surface reconstruction against HSR.
****
Non-Contrastive Unsupervised Learning of Physiological Signals From Video
Jeremy Speth, Nathan Vance, Patrick Flynn, Adam Czajka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14464-14474
Subtle periodic signals such as blood volume pulse and respiration can be extracted from RGB video, enabling noncontact health monitoring at low cost. Advancements in remote pulse estimation -- or remote photoplethysmography (rPPG) -- are currently driven by deep learning solutions. However, modern approaches are trained and evaluated on benchmark datasets with ground truth from contact-PPG sensors. We present the first non-contrastive unsupervised learning framework for signal regression to mitigate the need for labelled video data. With minimal assumptions of periodicity and finite bandwidth, our approach discovers the blood volume pulse directly from unlabelled videos. We find that encouraging sparse power spectra within normal physiological bandlimits and variance over batches of power spectra is sufficient for learning visual features of periodic signals. We perform the first experiments utilizing unlabelled video data not specifically created for rPPG to train robust pulse rate estimators. Given the limited inductive biases and impressive empirical results, the approach is theoretically capable of discovering other periodic signals from video, enabling multiple physiological measurements without the need for ground truth signals.
****
FashionSAP: Symbols and Attributes Prompt for Fine-Grained Fashion Vision-Language Pre-Training
Yunpeng Han, Lisai Zhang, Qingcai Chen, Zhijian Chen, Zhonghua Li, Jianxin Yang, Zhao Cao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15028-15038
Fashion vision-language pre-training models have shown efficacy for a wide range of downstream tasks. However, general vision-language pre-training models pay less attention to fine-grained domain features, while these features are important in distinguishing the specific domain tasks from general tasks. We propose a method for fine-grained fashion vision-language pre-training based on fashion Symbols and Attributes Prompt (FashionSAP) to model fine-grained multi-modalities fashion attributes and characteristics. Firstly, we propose the fashion symbols, a novel abstract fashion concept layer, to represent different fashion items and to generalize various kinds of fine-grained fashion features, making modelling fine-grained attributes more effective. Secondly, the attributes prompt method is proposed to make the model learn specific attributes of fashion items explicitly. We design proper prompt templates according to the format of fashion data. Comprehensive experiments are conducted on two public fashion benchmarks, i.e., F

ashionGen and FashionIQ, and FashionSAP gets SOTA performances for four popular fashion tasks. The ablation study also shows the proposed abstract fashion symbo ls, and the attribute prompt method enables the model to acquire fine-grained se mantics in the fashion domain effectively. The obvious performance gains from Fa shionSAP provide a new baseline for future fashion task research.
****
PartSLIP: Low-Shot Part Segmentation for 3D Point Clouds via Pretrained Image-La nguage Models
Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, Hao S u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2023, pp. 21736-21746
Generalizable 3D part segmentation is important but challenging in vision and ro botics. Training deep models via conventional supervised methods requires large-scale 3D datasets with fine-grained part annotations, which are costly to collec t. This paper explores an alternative way for low-shot part segmentation of 3D p oint clouds by leveraging a pretrained image-language model, GLIP, which achieve s superior performance on open-vocabulary 2D detection. We transfer the rich kno wledge from 2D to 3D through GLIP-based part detection on point cloud rendering and a novel 2D-to-3D label lifting algorithm. We also utilize multi-view 3D prio rs and few-shot prompt tuning to boost performance significantly. Extensive eval uation on PartNet and PartNet-Mobility datasets shows that our method enables ex cellent zero-shot 3D part segmentation. Our few-shot version not only outperform s existing few-shot approaches by a large margin but also achieves highly compet itive results compared to the fully supervised counterpart. Furthermore, we demo nstrate that our method can be directly applied to iPhone-scanned point clouds w ithout significant domain gaps.
****
An Erudite Fine-Grained Visual Classification Model
Dongliang Chang, Yujun Tong, Ruoyi Du, Timothy Hospedales, Yi-Zhe Song, Zhanyu M a; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2023, pp. 7268-7277
Current fine-grained visual classification (FGVC) models are isolated. In practi ce, we first need to identify the coarse-grained label of an object, then select the corresponding FGVC model for recognition. This hinders the application of t he FGVC algorithm in real-life scenarios. In this paper, we propose an erudite F GVC model jointly trained by several different datasets, which can efficiently a nd accurately predict an object's fine-grained label across the combined label s pace. We found through a pilot study that positive and negative transfers co-occ ur when different datasets are mixed for training, i.e., the knowledge from othe r datasets is not always useful. Therefore, we first propose a feature disentang lement module and a feature re-fusion module to reduce negative transfer and boo st positive transfer between different datasets. In detail, we reduce negative t ransfer by decoupling the deep features through many dataset-specific feature ex tractors. Subsequently, these are channel-wise re-fused to facilitate positive t ransfer. Finally, we propose a meta-learning based dataset-agnostic spatial atte ntion layer to take full advantage of the multi-dataset training data, given tha t localisation is dataset-agnostic between different datasets. Experimental resu lts across 11 different mixed-datasets built on four different FGVC datasets dem onstrate the effectiveness of the proposed method. Furthermore, the proposed met hod can be easily combined with existing FGVC methods to obtain state-of-the-art results.
****
MAGVLT: Masked Generative Vision-and-Language Transformer
Sungwoong Kim, Daejin Jo, Donghoon Lee, Jongmin Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23338-2 3348
While generative modeling on multimodal image-text data has been actively develo ped with large-scale paired datasets, there have been limited attempts to genera te both image and text data by a single model rather than a generation of one fi xed modality conditioned on the other modality. In this paper, we explore a unif

ied generative vision-and-language (VL) model that can produce both images and text sequences. Especially, we propose a generative VL transformer based on the non-autoregressive mask prediction, named MAGVLT, and compare it with an autoregressive generative VL transformer (ARGVLT). In comparison to ARGVLT, the proposed MAGVLT enables bidirectional context encoding, fast decoding by parallel token predictions in an iterative refinement, and extended editing capabilities such as image and text infilling. For rigorous training of our MAGVLT with image-text pairs from scratch, we combine the image-to-text, text-to image, and joint image-and-text mask prediction tasks. Moreover, we devise two additional tasks based on the step-unrolled mask prediction and the selective prediction on the mixture of two image-text pairs. Experimental results on various downstream generation tasks of VL benchmarks show that our MAGVLT outperforms ARGVLT by a large margin even with significant inference speedup. Particularly, MAGVLT achieves competitive results on both zero-shot image-to-text and text-to-image generation tasks from MS-COCO by one moderate-sized model (fewer than 500M parameters) even without the use of monomodal data and networks.
****
Structure Aggregation for Cross-Spectral Stereo Image Guided Denoising
Zehua Sheng, Zhu Yu, Xiongwei Liu, Si-Yuan Cao, Yuqi Liu, Hui-Liang Shen, Huaqi Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13997-14006
To obtain clean images with salient structures from noisy observations, a growing trend in current denoising studies is to seek the help of additional guidance images with high signal-to-noise ratios, which are often acquired in different spectral bands such as near infrared. Although previous guided denoising methods basically require the input images to be well-aligned, a more common way to capture the paired noisy target and guidance images is to exploit a stereo camera system. However, current studies on cross-spectral stereo matching cannot fully guarantee the pixel-level registration accuracy, and rarely consider the case of noise contamination. In this work, for the first time, we propose a guided denoising framework for cross-spectral stereo images. Instead of aligning the input images via conventional stereo matching, we aggregate structures from the guidance image to estimate a clean structure map for the noisy target image, which is then used to regress the final denoising result with a spatially variant linear representation model. Based on this, we design a neural network, called as SANet, to complete the entire guided denoising process. Experimental results show that, our SANet can effectively transfer structures from an unaligned guidance image to the restoration result, and outperforms state-of-the-art denoisers on various stereo image datasets. Besides, our structure aggregation strategy also shows its potential to handle other unaligned guided restoration tasks such as super-resolution and deblurring. The source code is available at https://github.com/lustrouselixir/SANet.
****
Decoupling Human and Camera Motion From Videos in the Wild
Vickie Ye, Georgios Pavlakos, Jitendra Malik, Angjoo Kanazawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21222-21232
We propose a method to reconstruct global human trajectories from videos in the wild. Our optimization method decouples the camera and human motion, which allows us to place people in the same world coordinate frame. Most existing methods do not model the camera motion; methods that rely on the background pixels to infer 3D human motion usually require a full scene reconstruction, which is often not possible for in-the-wild videos. However, even when existing SLAM systems cannot recover accurate scene reconstructions, the background pixel motion still provides enough signal to constrain the camera motion. We show that relative camera estimates along with data-driven human motion priors can resolve the scene scale ambiguity and recover global human trajectories. Our method robustly recovers the global 3D trajectories of people in challenging in-the-wild videos, such as PoseTrack. We quantify our improvement over existing methods on 3D human dataset Egobody. We further demonstrate that our recovered camera scale allows us to r

eason about motion of multiple people in a shared coordinate frame, which improves performance of downstream tracking in PoseTrack. Code and additional results can be found at https://vye16.github.io/slahmr/.
****
DetCLIPv2: Scalable Open-Vocabulary Object Detection Pre-Training via Word-Region Alignment
Lewei Yao, Jianhua Han, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Hang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23497-23506
This paper presents DetCLIPv2, an efficient and scalable training framework that incorporates large-scale image-text pairs to achieve open-vocabulary object detection (OVD). Unlike previous OVD frameworks that typically rely on a pre-trained vision-language model (e.g., CLIP) or exploit image-text pairs via a pseudo labeling process, DetCLIPv2 directly learns the fine-grained word-region alignment from massive image-text pairs in an end-to-end manner. To accomplish this, we employ a maximum word-region similarity between region proposals and textual words to guide the contrastive objective. To enable the model to gain localization capability while learning broad concepts, DetCLIPv2 is trained with a hybrid supervision from detection, grounding and image-text pair data under a unified data formulation. By jointly training with an alternating scheme and adopting low-resolution input for image-text pairs, DetCLIPv2 exploits image-text pair data efficiently and effectively: DetCLIPv2 utilizes 13x more image-text pairs than DetCLIP with a similar training time and improves performance. With 13M image-text pairs for pre-training, DetCLIPv2 demonstrates superior open-vocabulary detection performance, e.g., DetCLIPv2 with Swin-T backbone achieves 40.4% zero-shot AP on the LVIS benchmark, which outperforms previous works GLIP/GLIPv2/DetCLIP by 14.4/11.4/4.5% AP, respectively, and even beats its fully-supervised counterpart by a large margin.
****
Adversarially Robust Neural Architecture Search for Graph Neural Networks
Beini Xie, Heng Chang, Ziwei Zhang, Xin Wang, Daixin Wang, Zhiqiang Zhang, Rex Ying, Wenwu Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8143-8152
Graph Neural Networks (GNNs) obtain tremendous success in modeling relational data. Still, they are prone to adversarial attacks, which are massive threats to applying GNNs to risk-sensitive domains. Existing defensive methods neither guarantee performance facing new data/tasks or adversarial attacks nor provide insights to understand GNN robustness from an architectural perspective. Neural Architecture Search (NAS) has the potential to solve this problem by automating GNN architecture designs. Nevertheless, current graph NAS approaches lack robust design and are vulnerable to adversarial attacks. To tackle these challenges, we propose a novel Robust Neural Architecture search framework for GNNs (G-RNA). Specifically, we design a robust search space for the message-passing mechanism by adding graph structure mask operations into the search space, which comprises various defensive operation candidates and allows us to search for defensive GNNs. Furthermore, we define a robustness metric to guide the search procedure, which helps to filter robust architectures. In this way, G-RNA helps understand GNN robustness from an architectural perspective and effectively searches for optimal adversarial robust GNNs. Extensive experimental results on benchmark datasets show that G-RNA significantly outperforms manually designed robust GNNs and vanilla graph NAS baselines by 12.1% to 23.4% under adversarial attacks.
****
Affordance Grounding From Demonstration Video To Target Image
Joya Chen, Difei Gao, Kevin Qinghong Lin, Mike Zheng Shou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6799-6808
Humans excel at learning from expert demonstrations and solving their own problems. To equip intelligent robots and assistants, such as AR glasses, with this ability, it is essential to ground human hand interactions (i.e., affordances) from demonstration videos and apply them to a target image like a user's AR glass v

iew. The video-to-image affordance grounding task is challenging due to (1) the need to predict fine-grained affordances, and (2) the limited training data, which inadequately covers video-image discrepancies and negatively impacts grounding. To tackle them, we propose Affordance Transformer (Afformer), which has a fine-grained transformer-based decoder that gradually refines affordance grounding. Moreover, we introduce Mask Affordance Hand (MaskAHand), a self-supervised pretraining technique for synthesizing video-image data and simulating context changes, enhancing affordance grounding across video-image discrepancies. Afformer with MaskAHand pre-training achieves state-of-the-art performance on multiple benchmarks, including a substantial 37% improvement on the OPRA dataset. Code is made available at https://github.com/showlab/afformer.
****
GrowSP: Unsupervised Semantic Segmentation of 3D Point Clouds
Zihui Zhang, Bo Yang, Bing Wang, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17619-17629
We study the problem of 3D semantic segmentation from raw point clouds. Unlike existing methods which primarily rely on a large amount of human annotations for training neural networks, we propose the first purely unsupervised method, called GrowSP, to successfully identify complex semantic classes for every point in 3D scenes, without needing any type of human labels or pretrained models. The key to our approach is to discover 3D semantic elements via progressive growing of superpoints. Our method consists of three major components, 1) the feature extractor to learn per-point features from input point clouds, 2) the superpoint constructor to progressively grow the sizes of superpoints, and 3) the semantic primitive clustering module to group superpoints into semantic elements for the final semantic segmentation. We extensively evaluate our method on multiple datasets, demonstrating superior performance over all unsupervised baselines and approaching the classic fully supervised PointNet. We hope our work could inspire more advanced methods for unsupervised 3D semantic learning.
****
RONO: Robust Discriminative Learning With Noisy Labels for 2D-3D Cross-Modal Retrieval
Yanglin Feng, Hongyuan Zhu, Dezhong Peng, Xi Peng, Peng Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11610-11619
Recently, with the advent of Metaverse and AI Generated Content, cross-modal retrieval becomes popular with a burst of 2D and 3D data. However, this problem is challenging given the heterogeneous structure and semantic discrepancies. Moreover, imperfect annotations are ubiquitous given the ambiguous 2D and 3D content, thus inevitably producing noisy labels to degrade the learning performance. To tackle the problem, this paper proposes a robust 2D-3D retrieval framework (RONO) to robustly learn from noisy multimodal data. Specifically, one novel Robust Discriminative Center Learning mechanism (RDCL) is proposed in RONO to adaptively distinguish clean and noisy samples for respectively providing them with positive and negative optimization directions, thus mitigating the negative impact of noisy labels. Besides, we present a Shared Space Consistency Learning mechanism (SSCL) to capture the intrinsic information inside the noisy data by minimizing the cross-modal and semantic discrepancy between common space and label space simultaneously. Comprehensive mathematical analyses are given to theoretically prove the noise tolerance of the proposed method. Furthermore, we conduct extensive experiments on four 3D-model multimodal datasets to verify the effectiveness of our method by comparing it with 15 state-of-the-art methods. Code is available at https://github.com/penghu-cs/RONO.
****
One-Stage 3D Whole-Body Mesh Recovery With Component Aware Transformer
Jing Lin, Ailing Zeng, Haoqian Wang, Lei Zhang, Yu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21159-21168
Whole-body mesh recovery aims to estimate the 3D human body, face, and hands parameters from a single image. It is challenging to perform this task with a singl

e network due to resolution issues, i.e., the face and hands are usually located in extremely small regions. Existing works usually detect hands and faces, enlarge their resolution to feed in a specific network to predict the parameter, and finally fuse the results. While this copy-paste pipeline can capture the fine-grained details of the face and hands, the connections between different parts cannot be easily recovered in late fusion, leading to implausible 3D rotation and unnatural pose. In this work, we propose a one-stage pipeline for expressive whole-body mesh recovery, named OSX, without separate networks for each part. Specifically, we design a Component Aware Transformer (CAT) composed of a global body encoder and a local face/hand decoder. The encoder predicts the body parameters and provides a high-quality feature map for the decoder, which performs a feature-level upsample-crop scheme to extract high-resolution part-specific features and adopt keypoint-guided deformable attention to estimate hand and face precisely. The whole pipeline is simple yet effective without any manual post-processing and naturally avoids implausible prediction. Comprehensive experiments demonstrate the effectiveness of OSX. Lastly, we build a large-scale Upper-Body dataset (UBody) with high-quality 2D and 3D whole-body annotations. It contains persons with partially visible bodies in diverse real-life scenarios to bridge the gap between the basic task and downstream applications.
****
Masked Jigsaw Puzzle: A Versatile Position Embedding for Vision Transformers
Bin Ren, Yahui Liu, Yue Song, Wei Bi, Rita Cucchiara, Nicu Sebe, Wei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20382-20391
Position Embeddings (PEs), an arguably indispensable component in Vision Transformers (ViTs), have been shown to improve the performance of ViTs on many vision tasks. However, PEs have a potentially high risk of privacy leakage since the spatial information of the input patches is exposed. This caveat naturally raises a series of interesting questions about the impact of PEs on accuracy, privacy, prediction consistency, etc. To tackle these issues, we propose a Masked Jigsaw Puzzle (MJP) position embedding method. In particular, MJP first shuffles the selected patches via our block-wise random jigsaw puzzle shuffle algorithm, and their corresponding PEs are occluded. Meanwhile, for the non-occluded patches, the PEs remain the original ones but their spatial relation is strengthened via our dense absolute localization regressor. The experimental results reveal that 1) PEs explicitly encode the 2D spatial relationship and lead to severe privacy leakage problems under gradient inversion attack; 2) Training ViTs with the naively shuffled patches can alleviate the problem, but it harms the accuracy; 3) Under a certain shuffle ratio, the proposed MJP not only boosts the performance and robustness on large-scale datasets (i.e., ImageNet-1K and ImageNet-C, -A/O) but also improves the privacy preservation ability under typical gradient attacks by a large margin. The source code and trained models are available at https://github.com/yhlleo/MJP.
****
LayoutDiffusion: Controllable Diffusion Model for Layout-to-Image Generation
Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22490-22499
Recently, diffusion models have achieved great success in image synthesis. However, when it comes to the layout-to-image generation where an image often has a complex scene of multiple objects, how to make strong control over both the global layout map and each detailed object remains a challenging task. In this paper, we propose a diffusion model named LayoutDiffusion that can obtain higher generation quality and greater controllability than the previous works. To overcome the difficult multimodal fusion of image and layout, we propose to construct a structural image patch with region information and transform the patched image into a special layout to fuse with the normal layout in a unified form. Moreover, Layout Fusion Module (LFM) and Object-aware Cross Attention (OaCA) are proposed to model the relationship among multiple objects and designed to be object-aware and position-sensitive, allowing for precisely controlling the spatial related i

nformation. Extensive experiments show that our LayoutDiffusion outperforms the previous SOTA methods on FID, CAS by relatively 46.35%, 26.70% on COCO-stuff and 44.29%, 41.82% on VG. Code is available at https://github.com/ZGCTroy/LayoutDif fusion.
****
DeepMAD: Mathematical Architecture Design for Deep Convolutional Neural Network
Xuan Shen, Yaohua Wang, Ming Lin, Yilun Huang, Hao Tang, Xiuyu Sun, Yanzhi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2023, pp. 6163-6173
The rapid advances in Vision Transformer (ViT) refresh the state-of-the-art perf ormances in various vision tasks, overshadowing the conventional CNN-based model s. This ignites a few recent striking-back research in the CNN world showing tha t pure CNN models can achieve as good performance as ViT models when carefully t uned. While encouraging, designing such high-performance CNN models is challengi ng, requiring non-trivial prior knowledge of network design. To this end, a nove l framework termed Mathematical Architecture Design for Deep CNN (DeepMAD) is pr oposed to design high-performance CNN models in a principled way. In DeepMAD, a CNN network is modeled as an information processing system whose expressiveness and effectiveness can be analytically formulated by their structural parameters. Then a constrained mathematical programming (MP) problem is proposed to optimiz e these structural parameters. The MP problem can be easily solved by off-the-sh elf MP solvers on CPUs with a small memory footprint. In addition, DeepMAD is a pure mathematical framework: no GPU or training data is required during network design. The superiority of DeepMAD is validated on multiple large-scale computer vision benchmark datasets. Notably on ImageNet-1k, only using conventional conv olutional layers, DeepMAD achieves 0.7% and 1.5% higher top-1 accuracy than Conv NeXt and Swin on Tiny level, and 0.8% and 0.9% higher on Small level.
****
DISC: Learning From Noisy Labels via Dynamic Instance-Specific Selection and Cor rection
Yifan Li, Hu Han, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24070-24079
Existing studies indicate that deep neural networks (DNNs) can eventually memori ze the label noise. We observe that the memorization strength of DNNs towards ea ch instance is different and can be represented by the confidence value, which b ecomes larger and larger during the training process. Based on this, we propose a Dynamic Instance-specific Selection and Correction method (DISC) for learning from noisy labels (LNL). We first use a two-view-based backbone for image classi fication, obtaining confidence for each image from two views. Then we propose a dynamic threshold strategy for each instance, based on the momentum of each inst ance's memorization strength in previous epochs to select and correct noisy labe led data. Benefiting from the dynamic threshold strategy and two-view learning, we can effectively group each instance into one of the three subsets (i.e., clea n, hard, and purified) based on the prediction consistency and discrepancy by tw o views at each epoch. Finally, we employ different regularization strategies to conquer subsets with different degrees of label noise, improving the whole netw ork's robustness. Comprehensive evaluations on three controllable and four real-world LNL benchmarks show that our method outperforms the state-of-the-art (SOTA) methods to leverage useful information in noisy data while alleviating the pol lution of label noise.
****
BBDM: Image-to-Image Translation With Brownian Bridge Diffusion Models
Bo Li, Kaitao Xue, Bin Liu, Yu-Kun Lai; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1952-1961
Image-to-image translation is an important and challenging problem in computer v ision and image processing. Diffusion models(DM) have shown great potentials for high-quality image synthesis, and have gained competitive performance on the ta sk of image-to-image translation. However, most of the existing diffusion models treat image-to-image translation as conditional generation processes, and suffe r heavily from the gap between distinct domains. In this paper, a novel image-to

-image translation method based on the Brownian Bridge Diffusion Model(BBDM) is proposed, which models image-to-image translation as a stochastic Brownian Bridge process, and learns the translation between two domains directly through the bidirectional diffusion process rather than a conditional generation process. To the best of our knowledge, it is the first work that proposes Brownian Bridge diffusion process for image-to-image translation. Experimental results on various benchmarks demonstrate that the proposed BBDM model achieves competitive performance through both visual inspection and measurable metrics.
****
ConQueR: Query Contrast Voxel-DETR for 3D Object Detection
Benjin Zhu, Zhe Wang, Shaoshuai Shi, Hang Xu, Lanqing Hong, Hongsheng Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9296-9305
Although DETR-based 3D detectors simplify the detection pipeline and achieve direct sparse predictions, their performance still lags behind dense detectors with post-processing for 3D object detection from point clouds. DETRs usually adopt a larger number of queries than GTs (e.g., 300 queries v.s. 40 objects in Waymo) in a scene, which inevitably incur many false positives during inference. In this paper, we propose a simple yet effective sparse 3D detector, named Query Contrast Voxel-DETR (ConQueR), to eliminate the challenging false positives, and achieve more accurate and sparser predictions. We observe that most false positives are highly overlapping in local regions, caused by the lack of explicit supervision to discriminate locally similar queries. We thus propose a Query Contrast mechanism to explicitly enhance queries towards their best-matched GTs over all unmatched query predictions. This is achieved by the construction of positive and negative GT-query pairs for each GT, and a contrastive loss to enhance positive GT-query pairs against negative ones based on feature similarities. ConQueR closes the gap of sparse and dense 3D detectors, and reduces 60% false positives. Our single-frame ConQueR achieves 71.6 mAPH/L2 on the challenging Waymo Open Dataset validation set, outperforming previous sota methods by over 2.0 mAPH/L2. Code: https://github.com/poodarchu/EFG.
****
Probing Neural Representations of Scene Perception in a Hippocampally Dependent Task Using Artificial Neural Networks
Markus Frey, Christian F. Doeller, Caswell Barry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2113-2121
Deep artificial neural networks (DNNs) trained through backpropagation provide effective models of the mammalian visual system, accurately capturing the hierarchy of neural responses through primary visual cortex to inferior temporal cortex (IT). However, the ability of these networks to explain representations in higher cortical areas is relatively lacking and considerably less well researched. For example, DNNs have been less successful as a model of the egocentric to allocentric transformation embodied by circuits in retrosplenial and posterior parietal cortex. We describe a novel scene perception benchmark inspired by a hippocampal dependent task, designed to probe the ability of DNNs to transform scenes viewed from different egocentric perspectives. Using a network architecture inspired by the connectivity between temporal lobe structures and the hippocampus, we demonstrate that DNNs trained using a triplet loss can learn this task. Moreover, by enforcing a factorized latent space, we can split information propagation into "what" and "where" pathways, which we use to reconstruct the input. This allows us to beat the state-of-the-art for unsupervised object segmentation on the CATER and MOVi-A,B,C benchmarks.
****
Imagen Editor and EditBench: Advancing and Evaluating Text-Guided Image Inpainting
Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, William Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18359-18369

Text-guided image editing can have a transformative impact in supporting creativ
e applications. A key challenge is to generate edits that are faithful to the in
put text prompt, while consistent with the input image. We present Imagen Editor
, a cascaded diffusion model, built by fine-tuning Imagen on text-guided image i
npainting. Imagen Editor's edits are faithful to the text prompts, which is acco
mplished by incorporating object detectors for proposing inpainting masks during
 training. In addition, text-guided image inpainting captures fine details in th
e input image by conditioning the cascaded pipeline on the original high resolut
ion image. To improve qualitative and quantitative evaluation, we introduce Edit
Bench, a systematic benchmark for text-guided image inpainting. EditBench evalua
tes inpainting edits on natural and generated images exploring objects, attribut
es, and scenes. Through extensive human evaluation on EditBench, we find that ob
ject-masking during training leads to across-the-board improvements in text-imag
e alignment -- such that Imagen Editor is preferred over DALL-E 2 and Stable Dif
fusion -- and, as a cohort, these models are better at object-rendering than tex
t-rendering, and handle material/color/size attributes better than count/shape a
ttributes.
****
Robust Multiview Point Cloud Registration With Reliable Pose Graph Initializatio
n and History Reweighting
Haiping Wang, Yuan Liu, Zhen Dong, Yulan Guo, Yu-Shen Liu, Wenping Wang, Bisheng
 Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2023, pp. 9506-9515
In this paper, we present a new method for the multiview registration of point c
loud. Previous multiview registration methods rely on exhaustive pairwise regist
ration to construct a densely-connected pose graph and apply Iteratively Reweigh
ted Least Square (IRLS) on the pose graph to compute the scan poses. However, co
nstructing a densely-connected graph is time-consuming and contains lots of outl
ier edges, which makes the subsequent IRLS struggle to find correct poses. To ad
dress the above problems, we first propose to use a neural network to estimate t
he overlap between scan pairs, which enables us to construct a sparse but reliab
le pose graph. Then, we design a novel history reweighting function in the IRLS
scheme, which has strong robustness to outlier edges on the graph. In comparison
 with existing multiview registration methods, our method achieves 11% higher re
gistration recall on the 3DMatch dataset and 13% lower registration errors on t
he ScanNet dataset while reducing 70% required pairwise registrations. Comprehe
nsive ablation studies are conducted to demonstrate the effectiveness of our des
igns. The source code is available at https://github.com/WHU-USI3DV/SGHR.
****
A Probabilistic Framework for Lifelong Test-Time Adaptation
Dhanajit Brahma, Piyush Rai; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2023, pp. 3582-3591
Test-time adaptation (TTA) is the problem of updating a pre-trained source model
 at inference time given test input(s) from a different target domain. Most exis
ting TTA approaches assume the setting in which the target domain is stationary,
 i.e., all the test inputs come from a single target domain. However, in many pr
actical settings, the test input distribution might exhibit a lifelong/continual
 shift over time. Moreover, existing TTA approaches also lack the ability to pro
vide reliable uncertainty estimates, which is crucial when distribution shifts o
ccur between the source and target domain. To address these issues, we present P
ETAL (Probabilistic lifElong Test-time Adaptation with seLf-training prior), whi
ch solves lifelong TTA using a probabilistic approach, and naturally results in
(1) a student-teacher framework, where the teacher model is an exponential movin
g average of the student model, and (2) regularizing the model updates at infere
nce time using the source model as a regularizer. To prevent model drift in the
lifelong/continual TTA setting, we also propose a data-driven parameter restorat
ion technique which contributes to reducing the error accumulation and maintaini
ng the knowledge of recent domains by restoring only the irrelevant parameters.
In terms of predictive error rate as well as uncertainty based metrics such as B
rier score and negative log-likelihood, our method achieves better results than

the current state-of-the-art for online lifelong test-time adaptation across various benchmarks, such as CIFAR-10C, CIFAR-100C, ImageNetC, and ImageNet3DCC datasets. The source code for our approach is accessible at https://github.com/dhanajitb/petal.
****
Sound to Visual Scene Generation by Audio-to-Visual Latent Alignment
Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, Tae-Hyun Oh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6430-6440
How does audio describe the world around us? In this paper, we propose a method for generating an image of a scene from sound. Our method addresses the challenges of dealing with the large gaps that often exist between sight and sound. We design a model that works by scheduling the learning procedure of each model component to associate audio-visual modalities despite their information gaps. The key idea is to enrich the audio features with visual information by learning to align audio to visual latent space. We translate the input audio to visual features, then use a pre-trained generator to produce an image. To further improve the quality of our generated images, we use sound source localization to select the audio-visual pairs that have strong cross-modal correlations. We obtain substantially better results on the VEGAS and VGGSound datasets than prior approaches. We also show that we can control our model's predictions by applying simple manipulations to the input waveform, or to the latent space.
****
OSRT: Omnidirectional Image Super-Resolution With Distortion-Aware Transformer
Fanghua Yu, Xintao Wang, Mingdeng Cao, Gen Li, Ying Shan, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13283-13292
Omnidirectional images (ODIs) have obtained lots of research interest for immersive experiences. Although ODIs require extremely high resolution to capture details of the entire scene, the resolutions of most ODIs are insufficient. Previous methods attempt to solve this issue by image super-resolution (SR) on equirectangular projection (ERP) images. However, they omit geometric properties of ERP in the degradation process, and their models can hardly generalize to real ERP images. In this paper, we propose Fisheye downsampling, which mimics the real-world imaging process and synthesizes more realistic low-resolution samples. Then we design a distortion-aware Transformer (OSRT) to modulate ERP distortions continuously and self-adaptively. Without a cumbersome process, OSRT outperforms previous methods by about 0.2dB on PSNR. Moreover, we propose a convenient data augmentation strategy, which synthesizes pseudo ERP images from plain images. This simple strategy can alleviate the over-fitting problem of large networks and significantly boost the performance of ODI SR. Extensive experiments have demonstrated the state-of-the-art performance of our OSRT.
****
Text With Knowledge Graph Augmented Transformer for Video Captioning
Xin Gu, Guang Chen, Yufei Wang, Libo Zhang, Tiejian Luo, Longyin Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18941-18951
Video captioning aims to describe the content of videos using natural language. Although significant progress has been made, there is still much room to improve the performance for real-world applications, mainly due to the long-tail and open set issues of words. In this paper, we propose a text with knowledge graph augmented transformer (TextKG) for video captioning. Notably, TextKG is a two-stream transformer, formed by the external stream and internal stream. The external stream is designed to absorb external knowledge, which models the interactions between the external knowledge, e.g., pre-built knowledge graph, and the built-in information of videos, e.g., the salient object regions, speech transcripts, and video captions, to mitigate the open set of words challenge. Meanwhile, the internal stream is designed to exploit the multi-modality information in original videos (e.g., the appearance of video frames, speech transcripts, and video captions) to deal with the long-tail issue. In addition, the cross attention mechani

sm is also used in both streams to share information. In this way, the two streams can help each other for more accurate results. Extensive experiments conducted on four challenging video captioning datasets, i.e., YouCookII, ActivityNet Captions, MSR-VTT, and MSVD, demonstrate that the proposed method performs favorably against the state-of-the-art methods. Specifically, the proposed TextKG method outperforms the best published results by improving 18.7% absolute CIDEr scores on the YouCookII dataset.
****
Filtering, Distillation, and Hard Negatives for Vision-Language Pre-Training
Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, Dhruv Mahajan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6967-6977
Vision-language models trained with contrastive learning on large-scale noisy data are becoming increasingly popular for zero-shot recognition problems. In this paper we improve the following three aspects of the contrastive pre-training pipeline: dataset noise, model initialization and the training objective. First, we propose a straightforward filtering strategy titled Complexity, Action, and Text-spotting (CAT) that significantly reduces dataset size, while achieving improved performance across zero-shot vision-language tasks. Next, we propose an approach titled Concept Distillation to leverage strong unimodal representations for contrastive training that does not increase training complexity while outperforming prior work. Finally, we modify the traditional contrastive alignment objective, and propose an importance-sampling approach to up-sample the importance of hard-negatives without adding additional complexity. On an extensive zero-shot benchmark of 29 tasks, our Distilled and Hard-negative Training (DiHT) approach improves on 20 tasks compared to the baseline. Furthermore, for few-shot linear probing, we propose a novel approach that bridges the gap between zero-shot and few-shot performance, substantially improving over prior work. Models are available at github.com/facebookresearch/diht.
****
PointCMP: Contrastive Mask Prediction for Self-Supervised Learning on Point Cloud Videos
Zhiqiang Shen, Xiaoxiao Sheng, Longguang Wang, Yulan Guo, Qiong Liu, Xi Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1212-1222
Self-supervised learning can extract representations of good quality from solely unlabeled data, which is appealing for point cloud videos due to their high labelling cost. In this paper, we propose a contrastive mask prediction (PointCMP) framework for self-supervised learning on point cloud videos. Specifically, our PointCMP employs a two-branch structure to achieve simultaneous learning of both local and global spatio-temporal information. On top of this two-branch structure, a mutual similarity based augmentation module is developed to synthesize hard samples at the feature level. By masking dominant tokens and erasing principal channels, we generate hard samples to facilitate learning representations with better discrimination and generalization performance. Extensive experiments show that our PointCMP achieves the state-of-the-art performance on benchmark datasets and outperforms existing full-supervised counterparts. Transfer learning results demonstrate the superiority of the learned representations across different datasets and tasks.
****
IS-GGT: Iterative Scene Graph Generation With Generative Transformers
Sanjoy Kundu, Sathyanarayanan N. Aakur; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6292-6301
Scene graphs provide a rich, structured representation of a scene by encoding the entities (objects) and their spatial relationships in a graphical format. This representation has proven useful in several tasks, such as question answering, captioning, and even object detection, to name a few. Current approaches take a generation-by-classification approach where the scene graph is generated through labeling of all possible edges between objects in a scene, which adds computati

onal overhead to the approach. This work introduces a generative transformer-based approach to generating scene graphs beyond link prediction. Using two transformer-based components, we first sample a possible scene graph structure from detected objects and their visual features. We then perform predicate classification on the sampled edges to generate the final scene graph. This approach allows us to efficiently generate scene graphs from images with minimal inference overhead. Extensive experiments on the Visual Genome dataset demonstrate the efficiency of the proposed approach. Without bells and whistles, we obtain, on average, 20.7% mean recall (mR@100) across different settings for scene graph generation (SGG), outperforming state-of-the-art SGG approaches while offering competitive performance to unbiased SGG approaches.
****
Meta Omnium: A Benchmark for General-Purpose Learning-To-Learn
Ondrej Bohdal, Yinbing Tian, Yongshuo Zong, Ruchika Chavhan, Da Li, Henry Gouk, Li Guo, Timothy Hospedales; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7693-7703
Meta-learning and other approaches to few-shot learning are widely studied for image recognition, and are increasingly applied to other vision tasks such as pose estimation and dense prediction. This naturally raises the question of whether there is any few-shot meta-learning algorithm capable of generalizing across these diverse task types? To support the community in answering this question, we introduce Meta Omnium, a dataset-of-datasets spanning multiple vision tasks including recognition, keypoint localization, semantic segmentation and regression. We experiment with popular few-shot meta-learning baselines and analyze their ability to generalize across tasks and to transfer knowledge between them. Meta Omnium enables meta-learning researchers to evaluate model generalization to a much wider array of tasks than previously possible, and provides a single framework for evaluating meta-learners across a wide suite of vision applications in a consistent manner.
****
Multimodal Industrial Anomaly Detection via Hybrid Fusion
Yue Wang, Jinlong Peng, Jiangning Zhang, Ran Yi, Yabiao Wang, Chengjie Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8032-8041
2D-based Industrial Anomaly Detection has been widely discussed, however, multimodal industrial anomaly detection based on 3D point clouds and RGB images still has many untouched fields. Existing multimodal industrial anomaly detection methods directly concatenate the multimodal features, which leads to a strong disturbance between features and harms the detection performance. In this paper, we propose Multi-3D-Memory (M3DM), a novel multimodal anomaly detection method with hybrid fusion scheme: firstly, we design an unsupervised feature fusion with patch-wise contrastive learning to encourage the interaction of different modal features; secondly, we use a decision layer fusion with multiple memory banks to avoid loss of information and additional novelty classifiers to make the final decision. We further propose a point feature alignment operation to better align the point cloud and RGB features. Extensive experiments show that our multimodal industrial anomaly detection model outperforms the state-of-the-art (SOTA) methods on both detection and segmentation precision on MVTec-3D AD dataset. Code at github.com/nomewang/M3DM.
****
BEV@DC: Bird's-Eye View Assisted Training for Depth Completion
Wending Zhou, Xu Yan, Yinghong Liao, Yuankai Lin, Jin Huang, Gangming Zhao, Shuguang Cui, Zhen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9233-9242
Depth completion plays a crucial role in autonomous driving, in which cameras and LiDARs are two complementary sensors. Recent approaches attempt to exploit spatial geometric constraints hidden in LiDARs to enhance image-guided depth completion. However, only low efficiency and poor generalization can be achieved. In this paper, we propose BEV@DC, a more efficient and powerful multi-modal training scheme, to boost the performance of image-guided depth completion. In practice,

the proposed BEV@DC model comprehensively takes advantage of LiDARs with rich geometric details in training, employing an enhanced depth completion manner in inference, which takes only images (RGB and depth) as input. Specifically, the geometric-aware LiDAR features are projected onto a unified BEV space, combining with RGB features to perform BEV completion. By equipping a newly proposed point-voxel spatial propagation network (PV-SPN), this auxiliary branch introduces strong guidance to the original image branches via 3D dense supervision and feature consistency. As a result, our baseline model demonstrates significant improvements with the sole image inputs. Concretely, it achieves state-of-the-art on several benchmarks, e.g., ranking Top-1 on the challenging KITTI depth completion benchmark.
****
BoxTeacher: Exploring High-Quality Pseudo Labels for Weakly Supervised Instance Segmentation
Tianheng Cheng, Xinggang Wang, Shaoyu Chen, Qian Zhang, Wenyu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3145-3154
Labeling objects with pixel-wise segmentation requires a huge amount of human labor compared to bounding boxes. Most existing methods for weakly supervised instance segmentation focus on designing heuristic losses with priors from bounding boxes. While, we find that box-supervised methods can produce some fine segmentation masks and we wonder whether the detectors could learn from these fine masks while ignoring low-quality masks. To answer this question, we present BoxTeacher, an efficient and end-to-end training framework for high-performance weakly supervised instance segmentation, which leverages a sophisticated teacher to generate high-quality masks as pseudo labels. Considering the massive noisy masks hurt the training, we present a mask-aware confidence score to estimate the quality of pseudo masks and propose the noise-aware pixel loss and noise-reduced affinity loss to adaptively optimize the student with pseudo masks. Extensive experiments can demonstrate the effectiveness of the proposed BoxTeacher. Without bells and whistles, BoxTeacher remarkably achieves 35.0 mask AP and 36.5 mask AP with ResNet-50 and ResNet-101 respectively on the challenging COCO dataset, which outperforms the previous state-of-the-art methods by a significant margin and bridges the gap between box-supervised and mask-supervised methods. The code and models will be available later.
****
Change-Aware Sampling and Contrastive Learning for Satellite Images
Utkarsh Mall, Bharath Hariharan, Kavita Bala; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5261-5270
Automatic remote sensing tools can help inform many large-scale challenges such as disaster management, climate change, etc. While a vast amount of spatio-temporal satellite image data is readily available, most of it remains unlabelled. Without labels, this data is not very useful for supervised learning algorithms. Self-supervised learning instead provides a way to learn effective representations for various downstream tasks without labels. In this work, we leverage characteristics unique to satellite images to learn better self-supervised features. Specifically, we use the temporal signal to contrast images with long-term and short-term differences, and we leverage the fact that satellite images do not change frequently. Using these characteristics, we formulate a new loss contrastive loss called Change-Aware Contrastive (CACo) Loss. Further, we also present a novel method of sampling different geographical regions. We show that leveraging these properties leads to better performance on diverse downstream tasks. For example, we see a 6.5% relative improvement for semantic segmentation and an 8.5% relative improvement for change detection over the best-performing baseline with our method.
****
Large-Scale Training Data Search for Object Re-Identification
Yue Yao, Tom Gedeon, Liang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15568-15578
We consider a scenario where we have access to the target domain, but cannot aff

ord on-the-fly training data annotation, and instead would like to construct an alternative training set from a large-scale data pool such that a competitive model can be obtained. We propose a search and pruning (SnP) solution to this training data search problem, tailored to object re-identification (re-ID), an application aiming to match the same object captured by different cameras. Specifically, the search stage identifies and merges clusters of source identities which exhibit similar distributions with the target domain. The second stage, subject to a budget, then selects identities and their images from the Stage I output, to control the size of the resulting training set for efficient training. The two steps provide us with training sets 80% smaller than the source pool while achieving a similar or even higher re-ID accuracy. These training sets are also shown to be superior to a few existing search methods such as random sampling and greedy sampling under the same budget on training data size. If we release the budget, training sets resulting from the first stage alone allow even higher re-ID accuracy. We provide interesting discussions on the specificity of our method to the re-ID problem and particularly its role in bridging the re-ID domain gap. The code is available at https://github.com/yorkeyao/SnP.
****
Devil Is in the Queries: Advancing Mask Transformers for Real-World Medical Image Segmentation and Out-of-Distribution Localization
Mingze Yuan, Yingda Xia, Hexin Dong, Zifan Chen, Jiawen Yao, Mingyan Qiu, Ke Yan, Xiaoli Yin, Yu Shi, Xin Chen, Zaiyi Liu, Bin Dong, Jingren Zhou, Le Lu, Ling Zhang, Li Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23879-23889
Real-world medical image segmentation has tremendous long-tailed complexity of objects, among which tail conditions correlate with relatively rare diseases and are clinically significant. A trustworthy medical AI algorithm should demonstrate its effectiveness on tail conditions to avoid clinically dangerous damage in these out-of-distribution (OOD) cases. In this paper, we adopt the concept of object queries in Mask transformers to formulate semantic segmentation as a soft cluster assignment. The queries fit the feature-level cluster centers of inliers during training. Therefore, when performing inference on a medical image in real-world scenarios, the similarity between pixels and the queries detects and localizes OOD regions. We term this OOD localization as MaxQuery. Furthermore, the foregrounds of real-world medical images, whether OOD objects or inliers, are lesions. The difference between them is obviously less than that between the foreground and background, resulting in the object queries may focus redundantly on the background. Thus, we propose a query-distribution (QD) loss to enforce clear boundaries between segmentation targets and other regions at the query level, improving the inlier segmentation and OOD indication. Our proposed framework is tested on two real-world segmentation tasks, i.e., segmentation of pancreatic and liver tumors, outperforming previous leading algorithms by an average of 7.39% on AUROC, 14.69% on AUPR, and 13.79% on FPR95 for OOD localization. On the other hand, our framework improves the performance of inlier segmentation by an average of 5.27% DSC compared with nnUNet.
****
KD-DLGAN: Data Limited Image Generation via Knowledge Distillation
Kaiwen Cui, Yingchen Yu, Fangneng Zhan, Shengcai Liao, Shijian Lu, Eric P. Xing; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3872-3882
Generative Adversarial Networks (GANs) rely heavily on large-scale training data for training high-quality image generation models. With limited training data, the GAN discriminator often suffers from severe overfitting which directly leads to degraded generation especially in generation diversity. Inspired by the recent advances in knowledge distillation (KD), we propose KD-GAN, a knowledge-distillation based generation framework that introduces pre-trained vision-language models for training effective data-limited image generation models. KD-GAN consists of two innovative designs. The first is aggregated generative KD that mitigates the discriminator overfitting by challenging the discriminator with harder learning tasks and distilling more generalizable knowledge from the pre-trained mo

dels. The second is correlated generative KD that improves the generation divers ity by distilling and preserving the diverse image-text correlation within the p re-trained models. Extensive experiments over multiple benchmarks show that KD-G AN achieves superior image generation with limited training data. In addition, K D-GAN complements the state-of-the-art with consistent and substantial performan ce gains. Note that codes will be released.
****
Batch Model Consolidation: A Multi-Task Model Consolidation Framework
Iordanis Fostiropoulos, Jiaye Zhu, Laurent Itti; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3664-3676
In Continual Learning (CL), a model is required to learn a stream of tasks seque ntially without significant performance degradation on previously learned tasks. Current approaches fail for a long sequence of tasks from diverse domains and d ifficulties. Many of the existing CL approaches are difficult to apply in practi ce due to excessive memory cost or training time, or are tightly coupled to a si ngle device. With the intuition derived from the widely applied mini-batch train ing, we propose Batch Model Consolidation (BMC) to support more realistic CL und er conditions where multiple agents are exposed to a range of tasks. During a re gularization phase, BMC trains multiple expert models in parallel on a set of di sjoint tasks. Each expert maintains weight similarity to a base model through a stability loss, and constructs a buffer from a fraction of the task's data. Duri ng the consolidation phase, we combine the learned knowledge on 'batches' of exp ert models using a batched consolidation loss in memory data that aggregates all buffers. We thoroughly evaluate each component of our method in an ablation stu dy and demonstrate the effectiveness on standardized benchmark datasets Split-CI FAR-100, Tiny-ImageNet, and the Stream dataset composed of 71 image classificati on tasks from diverse domains and difficulties. Our method outperforms the next best CL approach by 70% and is the only approach that can maintain performance a t the end of 71 tasks.
****
SelfME: Self-Supervised Motion Learning for Micro-Expression Recognition
Xinqi Fan, Xueli Chen, Mingjie Jiang, Ali Raza Shahid, Hong Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13834-13843
Facial micro-expressions (MEs) refer to brief spontaneous facial movements that can reveal a person's genuine emotion. They are valuable in lie detection, crimi nal analysis, and other areas. While deep learning-based ME recognition (MER) me thods achieved impressive success, these methods typically require pre-processin g using conventional optical flow-based methods to extract facial motions as inp uts. To overcome this limitation, we proposed a novel MER framework using self-s upervised learning to extract facial motion for ME (SelfME). To the best of our knowledge, this is the first work using an automatically self-learned motion tec hnique for MER. However, the self-supervised motion learning method might suffer from ignoring symmetrical facial actions on the left and right sides of faces w hen extracting fine features. To address this issue, we developed a symmetric co ntrastive vision transformer (SCViT) to constrain the learning of similar facial action features for the left and right parts of faces. Experiments were conduct ed on two benchmark datasets showing that our method achieved state-of-the-art p erformance, and ablation studies demonstrated the effectiveness of our method.
****
DR2: Diffusion-Based Robust Degradation Remover for Blind Face Restoration
Zhixin Wang, Ziying Zhang, Xiaoyun Zhang, Huangjie Zheng, Mingyuan Zhou, Ya Zhan g, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2023, pp. 1704-1713
Blind face restoration usually synthesizes degraded low-quality data with a pre-defined degradation model for training, while more complex cases could happen in the real world. This gap between the assumed and actual degradation hurts the r estoration performance where artifacts are often observed in the output. However , it is expensive and infeasible to include every type of degradation to cover r eal-world cases in the training data. To tackle this robustness issue, we propos

e Diffusion-based Robust Degradation Remover (DR2) to first transform the degraded image to a coarse but degradation-invariant prediction, then employ an enhancement module to restore the coarse prediction to a high-quality image. By leveraging a well-performing denoising diffusion probabilistic model, our DR2 diffuses input images to a noisy status where various types of degradation give way to Gaussian noise, and then captures semantic information through iterative denoising steps. As a result, DR2 is robust against common degradation (e.g. blur, resize, noise and compression) and compatible with different designs of enhancement modules. Experiments in various settings show that our framework outperforms state-of-the-art methods on heavily degraded synthetic and real-world datasets.
****
T-SEA: Transfer-Based Self-Ensemble Attack on Object Detection
Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, Kevin Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20514-20523
Compared to query-based black-box attacks, transfer-based black-box attacks do not require any information of the attacked models, which ensures their secrecy. However, most existing transfer-based approaches rely on ensembling multiple models to boost the attack transferability, which is time- and resource-intensive, not to mention the difficulty of obtaining diverse models on the same task. To address this limitation, in this work, we focus on the single-model transfer-based black-box attack on object detection, utilizing only one model to achieve a high-transferability adversarial attack on multiple black-box detectors. Specifically, we first make observations on the patch optimization process of the existing method and propose an enhanced attack framework by slightly adjusting its training strategies. Then, we analogize patch optimization with regular model optimization, proposing a series of self-ensemble approaches on the input data, the attacked model, and the adversarial patch to efficiently make use of the limited information and prevent the patch from overfitting. The experimental results show that the proposed framework can be applied with multiple classical base attack methods (e.g., PGD and MIM) to greatly improve the black-box transferability of the well-optimized patch on multiple mainstream detectors, meanwhile boosting white-box performance.
****
LiDAR2Map: In Defense of LiDAR-Based Semantic Map Construction Using Online Camera Distillation
Song Wang, Wentong Li, Wenyu Liu, Xiaolu Liu, Jianke Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5186-5195
Semantic map construction under bird's-eye view (BEV) plays an essential role in autonomous driving. In contrast to camera image, LiDAR provides the accurate 3D observations to project the captured 3D features onto BEV space inherently. However, the vanilla LiDAR-based BEV feature often contains many indefinite noises, where the spatial features have little texture and semantic cues. In this paper, we propose an effective LiDAR-based method to build semantic map. Specifically, we introduce a BEV pyramid feature decoder that learns the robust multi-scale BEV features for semantic map construction, which greatly boosts the accuracy of the LiDAR-based method. To mitigate the defects caused by lacking semantic cues in LiDAR data, we present an online Camera-to-LiDAR distillation scheme to facilitate the semantic learning from image to point cloud. Our distillation scheme consists of feature-level and logit-level distillation to absorb the semantic information from camera in BEV. The experimental results on challenging nuScenes dataset demonstrate the efficacy of our proposed LiDAR2Map on semantic map construction, which significantly outperforms the previous LiDAR-based methods over 27.9% mIoU and even performs better than the state-of-the-art camera-based approaches. Source code is available at: https://github.com/songw-zju/LiDAR2Map.
****
NewsNet: A Novel Dataset for Hierarchical Temporal Segmentation
Haoqian Wu, Keyu Chen, Haozhe Liu, Mingchen Zhuge, Bing Li, Ruizhi Qiao, Xiujun Shu, Bei Gan, Liangsheng Xu, Bo Ren, Mengmeng Xu, Wentian Zhang, Raghavendra Ram

achandra, Chia-Wen Lin, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10669-10680
Temporal video segmentation is the get-to-go automatic video analysis, which decomposes a long-form video into smaller components for the following-up understanding tasks. Recent works have studied several levels of granularity to segment a video, such as shot, event, and scene. Those segmentations can help compare the semantics in the corresponding scales, but lack a wider view of larger temporal spans, especially when the video is complex and structured. Therefore, we present two abstractive levels of temporal segmentations and study their hierarchy to the existing fine-grained levels. Accordingly, we collect NewsNet, the largest news video dataset consisting of 1,000 videos in over 900 hours, associated with several tasks for hierarchical temporal video segmentation. Each news video is a collection of stories on different topics, represented as aligned audio, visual, and textual data, along with extensive frame-wise annotations in four granularities. We assert that the study on NewsNet can advance the understanding of complex structured video and benefit more areas such as short-video creation, personalized advertisement, digital instruction, and education. Our dataset and code is publicly available at: https://github.com/NewsNet-Benchmark/NewsNet.
****
Token Contrast for Weakly-Supervised Semantic Segmentation
Lixiang Ru, Heliang Zheng, Yibing Zhan, Bo Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3093-3102
Weakly-Supervised Semantic Segmentation (WSSS) using image-level labels typically utilizes Class Activation Map (CAM) to generate the pseudo labels. Limited by the local structure perception of CNN, CAM usually cannot identify the integral object regions. Though the recent Vision Transformer (ViT) can remedy this flaw, we observe it also brings the over-smoothing issue, ie, the final patch tokens incline to be uniform. In this work, we propose Token Contrast (ToCo) to address this issue and further explore the virtue of ViT for WSSS. Firstly, motivated by the observation that intermediate layers in ViT can still retain semantic diversity, we designed a Patch Token Contrast module (PTC). PTC supervises the final patch tokens with the pseudo token relations derived from intermediate layers, allowing them to align the semantic regions and thus yield more accurate CAM. Secondly, to further differentiate the low-confidence regions in CAM, we devised a Class Token Contrast module (CTC) inspired by the fact that class tokens in ViT can capture high-level semantics. CTC facilitates the representation consistency between uncertain local regions and global objects by contrasting their class tokens. Experiments on the PASCAL VOC and MS COCO datasets show the proposed ToCo can remarkably surpass other single-stage competitors and achieve comparable performance with state-of-the-art multi-stage methods. Code is available at https://github.com/rulixiang/ToCo.
****
LightedDepth: Video Depth Estimation in Light of Limited Inference View Angles
Shengjie Zhu, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5003-5012
Video depth estimation infers the dense scene depth from immediate neighboring video frames. While recent works consider it a simplified structure-from-motion (SfM) problem, it still differs from the SfM in that significantly fewer view angels are available in inference. This setting, however, suits the mono-depth and optical flow estimation. This observation motivates us to decouple the video depth estimation into two components, a normalized pose estimation over a flowmap and a logged residual depth estimation over a mono-depth map. The two parts are unified with an efficient off-the-shelf scale alignment algorithm. Additionally, we stabilize the indoor two-view pose estimation by including additional projection constraints and ensuring sufficient camera translation. Though a two-view algorithm, we validate the benefit of the decoupling with the substantial performance improvement over multi-view iterative prior works on indoor and outdoor datasets. Codes and models are available at https://github.com/ShngJZ/LightedDepth.
****
Uncertainty-Aware Unsupervised Image Deblurring With Deep Residual Prior

Xiaole Tang, Xile Zhao, Jun Liu, Jianli Wang, Yuchun Miao, Tieyong Zeng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9883-9892
Non-blind deblurring methods achieve decent performance under the accurate blur kernel assumption. Since the kernel uncertainty (i.e. kernel error) is inevitable in practice, semi-blind deblurring is suggested to handle it by introducing the prior of the kernel (or induced) error. However, how to design a suitable prior for the kernel (or induced) error remains challenging. Hand-crafted prior, incorporating domain knowledge, generally performs well but may lead to poor performance when kernel (or induced) error is complex. Data-driven prior, which excessively depends on the diversity and abundance of training data, is vulnerable to out-of-distribution blurs and images. To address this challenge, we suggest a dataset-free deep residual prior for the kernel induced error (termed as residual) expressed by a customized untrained deep neural network, which allows us to flexibly adapt to different blurs and images in real scenarios. By organically integrating the respective strengths of deep priors and hand-crafted priors, we propose an unsupervised semi-blind deblurring model which recovers the latent image from the blurry image and inaccurate blur kernel. To tackle the formulated model, an efficient alternating minimization algorithm is developed. Extensive experiments demonstrate the favorable performance of the proposed method as compared to model-driven and data-driven methods in terms of image quality and the robustness to different types of kernel error.
****
HouseDiffusion: Vector Floorplan Generation via a Diffusion Model With Discrete and Continuous Denoising
Mohammad Amin Shabani, Sepidehsadat Hosseini, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5466-5475
The paper presents a novel approach for vector-floorplan generation via a diffusion model, which denoises 2D coordinates of room/door corners with two inference objectives: 1) a single-step noise as the continuous quantity to precisely invert the continuous forward process; and 2) the final 2D coordinate as the discrete quantity to establish geometric incident relationships such as parallelism, orthogonality, and corner-sharing. Our task is graph-conditioned floorplan generation, a common workflow in floorplan design. We represent a floorplan as 1D polygonal loops, each of which corresponds to a room or a door. Our diffusion model employs a Transformer architecture at the core, which controls the attention masks based on the input graph-constraint and directly generates vector-graphics floorplans via a discrete and continuous denoising process. We have evaluated our approach on RPLAN dataset. The proposed approach makes significant improvements in all the metrics against the state-of-the-art with significant margins, while being capable of generating non-Manhattan structures and controlling the exact number of corners per room. We will share all our code and models.
****
FedDM: Iterative Distribution Matching for Communication-Efficient Federated Learning
Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, Cho-Jui Hsieh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16323-16332
Federated learning (FL) has recently attracted increasing attention from academia and industry, with the ultimate goal of achieving collaborative training under privacy and communication constraints. Existing iterative model averaging based FL algorithms require a large number of communication rounds to obtain a well-performed model due to extremely unbalanced and non-i.i.d data partitioning among different clients. Thus, we propose FedDM to build the global training objective from multiple local surrogate functions, which enables the server to gain a more global view of the loss landscape. In detail, we construct synthetic sets of data on each client to locally match the loss landscape from original data through distribution matching. FedDM reduces communication rounds and improves model quality by transmitting more informative and smaller synthesized data compared w

ith unwieldy model weights. We conduct extensive experiments on three image classification datasets, and results show that our method can outperform other FL counterparts in terms of efficiency and model performance. Moreover, we demonstrate that FedDM can be adapted to preserve differential privacy with Gaussian mechanism and train a better model under the same privacy budget.
****
V2X-Seq: A Large-Scale Sequential Dataset for Vehicle-Infrastructure Cooperative Perception and Forecasting
Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, Juan Song, Jirui Yuan, Ping Luo, Zaiqing Nie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5486-5495
Utilizing infrastructure and vehicle-side information to track and forecast the behaviors of surrounding traffic participants can significantly improve decision-making and safety in autonomous driving. However, the lack of real-world sequential datasets limits research in this area. To address this issue, we introduce V2X-Seq, the first large-scale sequential V2X dataset, which includes data frames, trajectories, vector maps, and traffic lights captured from natural scenery. V2X-Seq comprises two parts: the sequential perception dataset, which includes more than 15,000 frames captured from 95 scenarios, and the trajectory forecasting dataset, which contains about 80,000 infrastructure-view scenarios, 80,000 vehicle-view scenarios, and 50,000 cooperative-view scenarios captured from 28 intersections' areas, covering 672 hours of data. Based on V2X-Seq, we introduce three new tasks for vehicle-infrastructure cooperative (VIC) autonomous driving: VIC3D Tracking, Online-VIC Forecasting, and Offline-VIC Forecasting. We also provide benchmarks for the introduced tasks. Find data, code, and more up-to-date information at https://github.com/AIR-THU/DAIR-V2X-Seq.
****
PSVT: End-to-End Multi-Person 3D Pose and Shape Estimation With Progressive Video Transformers
Zhongwei Qiu, Qiansheng Yang, Jian Wang, Haocheng Feng, Junyu Han, Errui Ding, Chang Xu, Dongmei Fu, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21254-21263
Existing methods of multi-person video 3D human Pose and Shape Estimation (PSE) typically adopt a two-stage strategy, which first detects human instances in each frame and then performs single-person PSE with temporal model. However, the global spatio-temporal context among spatial instances can not be captured. In this paper, we propose a new end-to-end multi-person 3D Pose and Shape estimation framework with progressive Video Transformer, termed PSVT. In PSVT, a spatio-temporal encoder (STE) captures the global feature dependencies among spatial objects. Then, spatio-temporal pose decoder (STPD) and shape decoder (STSD) capture the global dependencies between pose queries and feature tokens, shape queries and feature tokens, respectively. To handle the variances of objects as time proceeds, a novel scheme of progressive decoding is used to update pose and shape queries at each frame. Besides, we propose a novel pose-guided attention (PGA) for shape decoder to better predict shape parameters. The two components strengthen the decoder of PSVT to improve performance. Extensive experiments on the four datasets show that PSVT achieves stage-of-the-art results.
****
Bit-Shrinking: Limiting Instantaneous Sharpness for Improving Post-Training Quantization
Chen Lin, Bo Peng, Zheyang Li, Wenming Tan, Ye Ren, Jun Xiao, Shiliang Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16196-16205
Post-training quantization (PTQ) is an effective compression method to reduce the model size and computational cost. However, quantizing a model into a low-bit one, e.g., lower than 4, is difficult and often results in nonnegligible performance degradation. To address this, we investigate the loss landscapes of quantized networks with various bit-widths. We show that the network with more ragged loss surface, is more easily trapped into bad local minima, which mostly appears

in low-bit quantization. A deeper analysis indicates, the ragged surface is caus
ed by the injection of excessive quantization noise. To this end, we detach a sh
arpness term from the loss which reflects the impact of quantization noise. To s
mooth the rugged loss surface, we propose to limit the sharpness term small and
stable during optimization. Instead of directly optimizing the target bit networ
k, the bit-width of quantized network has a self-adapted shrinking scheduler in
continuous domain from high bit-width to the target by limiting the increasing s
harpness term within a proper range. It can be viewed as iteratively adding smal
l "instant" quantization noise and adjusting the network to eliminate its impact
. Widely experiments including classification and detection tasks demonstrate th
e effectiveness of the Bit-shrinking strategy in PTQ. On the Vision Transformer
models, our INT8 and INT6 models drop within 0.5% and 1.5% Top-1 accuracy, respe
ctively. On the traditional CNN networks, our INT4 quantized models drop within
1.3% and 3.5% Top-1 accuracy on ResNet18 and MobileNetV2 without fine-tuning, wh
ich achieves the state-of-the-art performance.
****
LSTFE-Net:Long Short-Term Feature Enhancement Network for Video Small Object Det
ection
Jinsheng Xiao, Yuanxu Wu, Yunhua Chen, Shurui Wang, Zhongyuan Wang, Jiayi Ma; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2023, pp. 14613-14622
Video small object detection is a difficult task due to the lack of object infor
mation. Recent methods focus on adding more temporal information to obtain more
potent high-level features, which often fail to specify the most vital informati
on for small objects, resulting in insufficient or inappropriate features. Since
information from frames at different positions contributes differently to small
objects, it is not ideal to assume that using one universal method will extract
proper features. We find that context information from the long-term frame and
temporal information from the short-term frame are two useful cues for video sma
ll object detection. To fully utilize these two cues, we propose a long short-te
rm feature enhancement network (LSTFE-Net) for video small object detection. Fir
st, we develop a plug-and-play spatio-temporal feature alignment module to creat
e temporal correspondences between the short-term and current frames. Then, we p
ropose a frame selection module to select the long-term frame that can provide t
he most additional context information. Finally, we propose a long short-term fe
ature aggregation module to fuse long short-term features. Compared to other sta
te-of-the-art methods, our LSTFE-Net achieves 4.4% absolute boosts in AP on the
FL-Drones dataset. More details can be found at https://github.com/xiaojs18/LSTF
E-Net.
****
MIC: Masked Image Consistency for Context-Enhanced Domain Adaptation
Lukas Hoyer, Dengxin Dai, Haoran Wang, Luc Van Gool; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11721-1
1732
In unsupervised domain adaptation (UDA), a model trained on source data (e.g. sy
nthetic) is adapted to target data (e.g. real-world) without access to target an
notation. Most previous UDA methods struggle with classes that have a similar vi
sual appearance on the target domain as no ground truth is available to learn th
e slight appearance differences. To address this problem, we propose a Masked Im
age Consistency (MIC) module to enhance UDA by learning spatial context relation
s of the target domain as additional clues for robust visual recognition. MIC en
forces the consistency between predictions of masked target images, where random
patches are withheld, and pseudo-labels that are generated based on the complet
e image by an exponential moving average teacher. To minimize the consistency lo
ss, the network has to learn to infer the predictions of the masked regions from
their context. Due to its simple and universal concept, MIC can be integrated i
nto various UDA methods across different visual recognition tasks such as image
classification, semantic segmentation, and object detection. MIC significantly i
mproves the state-of-the-art performance across the different recognition tasks
for synthetic-to-real, day-to-nighttime, and clear-to-adverse-weather UDA. For i

nstance, MIC achieves an unprecedented UDA performance of 75.9 mIoU and 92.8% on GTA-to-Cityscapes and VisDA-2017, respectively, which corresponds to an improvement of +2.1 and +3.0 percent points over the previous state of the art. The implementation is available at https://github.com/lhoyer/MIC.
****
Bridging the Gap Between Model Explanations in Partially Annotated Multi-Label Classification
Youngwook Kim, Jae Myung Kim, Jieun Jeong, Cordelia Schmid, Zeynep Akata, Jungwoo Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3408-3417
Due to the expensive costs of collecting labels in multi-label classification datasets, partially annotated multi-label classification has become an emerging field in computer vision. One baseline approach to this task is to assume unobserved labels as negative labels, but this assumption induces label noise as a form of false negative. To understand the negative impact caused by false negative labels, we study how these labels affect the model's explanation. We observe that the explanation of two models, trained with full and partial labels each, highlights similar regions but with different scaling, where the latter tends to have lower attribution scores. Based on these findings, we propose to boost the attribution scores of the model trained with partial labels to make its explanation resemble that of the model trained with full labels. Even with the conceptually simple approach, the multi-label classification performance improves by a large margin in three different datasets on a single positive label setting and one on a large-scale partial label setting. Code is available at https://github.com/youngwk/BridgeGapExplanationPAMC.
****
SkyEye: Self-Supervised Bird's-Eye-View Semantic Mapping Using Monocular Frontal View Images
Nikhil Gosala, Kürsat Petek, Paulo L. J. Drews-Jr, Wolfram Burgard, Abhinav Valada; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14901-14910
Bird's-Eye-View (BEV) semantic maps have become an essential component of automated driving pipelines due to the rich representation they provide for decision-making tasks. However, existing approaches for generating these maps still follow a fully supervised training paradigm and hence rely on large amounts of annotated BEV data. In this work, we address this limitation by proposing the first self-supervised approach for generating a BEV semantic map using a single monocular image from the frontal view (FV). During training, we overcome the need for BEV ground truth annotations by leveraging the more easily available FV semantic annotations of video sequences. Thus, we propose the SkyEye architecture that learns based on two modes of self-supervision, namely, implicit supervision and explicit supervision. Implicit supervision trains the model by enforcing spatial consistency of the scene over time based on FV semantic sequences, while explicit supervision exploits BEV pseudolabels generated from FV semantic annotations and self-supervised depth estimates. Extensive evaluations on the KITTI-360 dataset demonstrate that our self-supervised approach performs on par with the state-of-the-art fully supervised methods and achieves competitive results using only 1% of direct supervision in BEV compared to fully supervised approaches. Finally, we publicly release both our code and the BEV datasets generated from the KITTI-360 and Waymo datasets.
****
Unifying Vision, Text, and Layout for Universal Document Processing
Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, Mohit Bansal; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19254-19264
We propose Universal Document Processing (UDOP), a foundation Document AI model which unifies text, image, and layout modalities together with varied task formats, including document understanding and generation. UDOP leverages the spatial correlation between textual content and document image to model image, text, and layout modalities with one uniform representation. With a novel Vision-Text-Lay

out Transformer, UDOP unifies pretraining and multi-domain downstream tasks into a prompt-based sequence generation scheme. UDOP is pretrained on both large-scale unlabeled document corpora using innovative self-supervised objectives and diverse labeled data. UDOP also learns to generate document images from text and layout modalities via masked image reconstruction. To the best of our knowledge, this is the first time in the field of document AI that one model simultaneously achieves high-quality neural document editing and content customization. Our method sets the state-of-the-art on 8 Document AI tasks, e.g., document understanding and QA, across diverse data domains like finance reports, academic papers, and websites. UDOP ranks first on the leaderboard of the Document Understanding Benchmark.
****
SparsePose: Sparse-View Camera Pose Regression and Refinement
Samarth Sinha, Jason Y. Zhang, Andrea Tagliasacchi, Igor Gilitschenski, David B. Lindell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21349-21359
Camera pose estimation is a key step in standard 3D reconstruction pipelines that operates on a dense set of images of a single object or scene. However, methods for pose estimation often fail when there are only a few images available because they rely on the ability to robustly identify and match visual features between pairs of images. While these methods can work robustly with dense camera views, capturing a large set of images can be time consuming or impractical. Here, we propose Sparse-View Camera Pose Regression and Refinement (SparsePose) for recovering accurate camera poses given a sparse set of wide-baseline images (fewer than 10). The method learns to regress initial camera poses and then iteratively refine them after training on a large-scale dataset of objects (Co3D: Common Objects in 3D). SparsePose significantly outperforms conventional and learning-based baselines in recovering accurate camera rotations and translations. We also demonstrate our pipeline for high-fidelity 3D reconstruction using only 5-9 images of an object.
****
Learning Audio-Visual Source Localization via False Negative Aware Contrastive Learning
Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, Nick Barnes; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6420-6429
Self-supervised audio-visual source localization aims to locate sound-source objects in video frames without extra annotations. Recent methods often approach this goal with the help of contrastive learning, which assumes only the audio and visual contents from the same video are positive samples for each other. However, this assumption would suffer from false negative samples in real-world training. For example, for an audio sample, treating the frames from the same audio class as negative samples may mislead the model and therefore harm the learned representations (e.g., the audio of a siren wailing may reasonably correspond to the ambulances in multiple images). Based on this observation, we propose a new learning strategy named False Negative Aware Contrastive (FNAC) to mitigate the problem of misleading the training with such false negative samples. Specifically, we utilize the intra-modal similarities to identify potentially similar samples and construct corresponding adjacency matrices to guide contrastive learning. Further, we propose to strengthen the role of true negative samples by explicitly leveraging the visual features of sound sources to facilitate the differentiation of authentic sounding source regions. FNAC achieves state-of-the-art performances on Flickr-SoundNet, VGG-Sound, and AVSBench, which demonstrates the effectiveness of our method in mitigating the false negative issue. The code is available at https://github.com/OpenNLPLab/FNAC_AVL
****
VoxFormer: Sparse Voxel Transformer for Camera-Based 3D Semantic Scene Completion
Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M. Alvarez, Sanja Fidler, Chen Feng, Anima Anandkumar; Proceedings of the IEEE/CVF Conference on Com

puter Vision and Pattern Recognition (CVPR), 2023, pp. 9087-9098
Humans can easily imagine the complete 3D geometry of occluded objects and scene s. This appealing ability is vital for recognition and understanding. To enable such capability in AI systems, we propose VoxFormer, a Transformer-based semanti c scene completion framework that can output complete 3D volumetric semantics fr om only 2D images. Our framework adopts a two-stage design where we start from a sparse set of visible and occupied voxel queries from depth estimation, followe d by a densification stage that generates dense 3D voxels from the sparse ones. A key idea of this design is that the visual features on 2D images correspond on ly to the visible scene structures rather than the occluded or empty spaces. The refore, starting with the featurization and prediction of the visible structures is more reliable. Once we obtain the set of sparse queries, we apply a masked a utoencoder design to propagate the information to all the voxels by self-attenti on. Experiments on SemanticKITTI show that VoxFormer outperforms the state of th e art with a relative improvement of 20.0% in geometry and 18.1% in semantics an d reduces GPU memory during training to less than 16GB. Our code is available on https://github.com/NVlabs/VoxFormer.
****
Joint Video Multi-Frame Interpolation and Deblurring Under Unknown Exposure Time
Wei Shang, Dongwei Ren, Yi Yang, Hongzhi Zhang, Kede Ma, Wangmeng Zuo; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13935-13944
Natural videos captured by consumer cameras often suffer from low framerate and motion blur due to the combination of dynamic scene complexity, lens and sensor imperfection, and less than ideal exposure setting. As a result, computational m ethods that jointly perform video frame interpolation and deblurring begin to em erge with the unrealistic assumption that the exposure time is known and fixed. In this work, we aim ambitiously for a more realistic and challenging task - joi nt video multi-frame interpolation and deblurring under unknown exposure time. T oward this goal, we first adopt a variant of supervised contrastive learning to construct an exposure-aware representation from input blurred frames. We then tr ain two U-Nets for intra-motion and inter-motion analysis, respectively, adaptin g to the learned exposure representation via gain tuning. We finally build our v ideo reconstruction network upon the exposure and motion representation by progr essive exposure-adaptive convolution and motion refinement. Extensive experiment s on both simulated and real-world datasets show that our optimized method achie ves notable performance gains over the state-of-the-art on the joint video x8 in terpolation and deblurring task. Moreover, on the seemingly implausible x16 inte rpolation task, our method outperforms existing methods by more than 1.5 dB in t erms of PSNR.
****
Flow Supervision for Deformable NeRF
Chaoyang Wang, Lachlan Ewen MacDonald, László A. Jeni, Simon Lucey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 23, pp. 21128-21137
In this paper we present a new method for deformable NeRF that can directly use optical flow as supervision. We overcome the major challenge with respect to the computationally inefficiency of enforcing the flow constraints to the backward deformation field, used by deformable NeRFs. Specifically, we show that invertin g the backward deformation function is actually not needed for computing scene f lows between frames. This insight dramatically simplifies the problem, as one is no longer constrained to deformation functions that can be analytically inverte d. Instead, thanks to the weak assumptions required by our derivation based on t he inverse function theorem, our approach can be extended to a broad class of co mmonly used backward deformation field. We present results on monocular novel vi ew synthesis with rapid object motion, and demonstrate significant improvements over baselines without flow supervision.
****
MMG-Ego4D: Multimodal Generalization in Egocentric Action Recognition
Xinyu Gong, Sreyas Mohan, Naina Dhingra, Jean-Charles Bazin, Yilei Li, Zhangyang

Wang, Rakesh Ranjan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6481-6491
In this paper, we study a novel problem in egocentric action recognition, which we term as "Multimodal Generalization" (MMG). MMG aims to study how systems can generalize when data from certain modalities is limited or even completely missing. We thoroughly investigate MMG in the context of standard supervised action recognition and the more challenging few-shot setting for learning new action categories. MMG consists of two novel scenarios, designed to support security, and efficiency considerations in real-world applications: (1) missing modality generalization where some modalities that were present during the train time are missing during the inference time, and (2) cross-modal zero-shot generalization, where the modalities present during the inference time and the training time are disjoint. To enable this investigation, we construct a new dataset MMG-Ego4D containing data points with video, audio, and inertial motion sensor (IMU) modalities. Our dataset is derived from Ego4D dataset, but processed and thoroughly re-annotated by human experts to facilitate research in the MMG problem. We evaluate a diverse array of models on MMG-Ego4D and propose new methods with improved generalization ability. In particular, we introduce a new fusion module with modality dropout training, contrastive-based alignment training, and a novel cross-modal prototypical loss for better few-shot performance. We hope this study will serve as a benchmark and guide future research in multimodal generalization problems. The benchmark and code are available at https://github.com/facebookresearch/MMG_Ego4D
****
Zero-Shot Text-to-Parameter Translation for Game Character Auto-Creation
Rui Zhao, Wei Li, Zhipeng Hu, Lincheng Li, Zhengxia Zou, Zhenwei Shi, Changjie Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21013-21023
Recent popular Role-Playing Games (RPGs) saw the great success of character auto-creation systems. The bone-driven face model controlled by continuous parameters (like the position of bones) and discrete parameters (like the hairstyles) makes it possible for users to personalize and customize in-game characters. Previous in-game character auto-creation systems are mostly image-driven, where facial parameters are optimized so that the rendered character looks similar to the reference face photo. This paper proposes a novel text-to-parameter translation method (T2P) to achieve zero-shot text-driven game character auto-creation. With our method, users can create a vivid in-game character with arbitrary text description without using any reference photo or editing hundreds of parameters manually. In our method, taking the power of large-scale pre-trained multi-modal CLIP and neural rendering, T2P searches both continuous facial parameters and discrete facial parameters in a unified framework. Due to the discontinuous parameter representation, previous methods have difficulty in effectively learning discrete facial parameters. T2P, to our best knowledge, is the first method that can handle the optimization of both discrete and continuous parameters. Experimental results show that T2P can generate high-quality and vivid game characters with given text prompts. T2P outperforms other SOTA text-to-3D generation methods on both objective evaluations and subjective evaluations.
****
PIVOT: Prompting for Video Continual Learning
Andrés Villa, Juan León Alcázar, Motasem Alfarra, Kumail Alhamoud, Julio Hurtado, Fabian Caba Heilbron, Alvaro Soto, Bernard Ghanem; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24214-24223
Modern machine learning pipelines are limited due to data availability, storage quotas, privacy regulations, and expensive annotation processes. These constraints make it difficult or impossible to train and update large-scale models on such dynamic annotated sets. Continual learning directly approaches this problem, with the ultimate goal of devising methods where a deep neural network effectively learns relevant patterns for new (unseen) classes, without significantly altering its performance on previously learned ones. In this paper, we address the pr

oblem of continual learning for video data. We introduce PIVOT, a novel method that leverages extensive knowledge in pre-trained models from the image domain, thereby reducing the number of trainable parameters and the associated forgetting. Unlike previous methods, ours is the first approach that effectively uses prompting mechanisms for continual learning without any in-domain pre-training. Our experiments show that PIVOT improves state-of-the-art methods by a significant 27% on the 20-task ActivityNet setup.
****
Dual-Bridging With Adversarial Noise Generation for Domain Adaptive rPPG Estimation
Jingda Du, Si-Qi Liu, Bochao Zhang, Pong C. Yuen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10355-10364
The remote photoplethysmography (rPPG) technique can estimate pulse-related metrics (e.g. heart rate and respiratory rate) from facial videos and has a high potential for health monitoring. The latest deep rPPG methods can model in-distribution noise due to head motion, video compression, etc., and estimate high-quality rPPG signals under similar scenarios. However, deep rPPG models may not generalize well to the target test domain with unseen noise and distortions. In this paper, to improve the generalization ability of rPPG models, we propose a dual-bridging network to reduce the domain discrepancy by aligning intermediate domains and synthesizing the target noise in the source domain for better noise reduction. To comprehensively explore the target domain noise, we propose a novel adversarial noise generation in which the noise generator indirectly competes with the noise reducer. To further improve the robustness of the noise reducer, we propose hard noise pattern mining to encourage the generator to learn hard noise patterns contained in the target domain features. We evaluated the proposed method on three public datasets with different types of interferences. Under different cross-domain scenarios, the comprehensive results show the effectiveness of our method.
****
Panoptic Video Scene Graph Generation
Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18675-18685
Towards building comprehensive real-world visual perception systems, we propose and study a new problem called panoptic scene graph generation (PVSG). PVSG is related to the existing video scene graph generation (VidSGG) problem, which focuses on temporal interactions between humans and objects localized with bounding boxes in videos. However, the limitation of bounding boxes in detecting non-rigid objects and backgrounds often causes VidSGG systems to miss key details that are crucial for comprehensive video understanding. In contrast, PVSG requires nodes in scene graphs to be grounded by more precise, pixel-level segmentation masks, which facilitate holistic scene understanding. To advance research in this new area, we contribute a high-quality PVSG dataset, which consists of 400 videos (289 third-person + 111 egocentric videos) with totally 150K frames labeled with panoptic segmentation masks as well as fine, temporal scene graphs. We also provide a variety of baseline methods and share useful design practices for future work.
****
3D Video Object Detection With Learnable Object-Centric Global Optimization
Jiawei He, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5106-5115
We explore long-term temporal visual correspondence-based optimization for 3D video object detection in this work. Visual correspondence refers to one-to-one mappings for pixels across multiple images. Correspondence-based optimization is the cornerstone for 3D scene reconstruction but is less studied in 3D video object detection, because moving objects violate multi-view geometry constraints and

are treated as outliers during scene reconstruction. We address this issue by treating objects as first-class citizens during correspondence-based optimization. In this work, we propose BA-Det, an end-to-end optimizable object detector with object-centric temporal correspondence learning and featuremetric object bundle adjustment. Empirically, we verify the effectiveness and efficiency of BA-Det for multiple baseline 3D detectors under various setups. Our BA-Det achieves SOTA performance on the large-scale Waymo Open Dataset (WOD) with only marginal computation cost. Our code is available at https://github.com/jiaweihe1996/BA-Det.
****
Improving the Transferability of Adversarial Samples by Path-Augmented Method
Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, Michael R. Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8173-8182
Deep neural networks have achieved unprecedented success on diverse vision tasks. However, they are vulnerable to adversarial noise that is imperceptible to humans. This phenomenon negatively affects their deployment in real-world scenarios, especially security-related ones. To evaluate the robustness of a target model in practice, transfer-based attacks craft adversarial samples with a local model and have attracted increasing attention from researchers due to their high efficiency. The state-of-the-art transfer-based attacks are generally based on data augmentation, which typically augments multiple training images from a linear path when learning adversarial samples. However, such methods selected the image augmentation path heuristically and may augment images that are semantics-inconsistent with the target images, which harms the transferability of the generated adversarial samples. To overcome the pitfall, we propose the Path-Augmented Method (PAM). Specifically, PAM first constructs a candidate augmentation path pool. It then settles the employed augmentation paths during adversarial sample generation with greedy search. Furthermore, to avoid augmenting semantics-inconsistent images, we train a Semantics Predictor (SP) to constrain the length of the augmentation path. Extensive experiments confirm that PAM can achieve an improvement of over 4.8% on average compared with the state-of-the-art baselines in terms of the attack success rates.
****
Robust Mean Teacher for Continual and Gradual Test-Time Adaptation
Mario Döbler, Robert A. Marsden, Bin Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7704-7714
Since experiencing domain shifts during test-time is inevitable in practice, test-time adaption (TTA) continues to adapt the model after deployment. Recently, the area of continual and gradual test-time adaptation (TTA) emerged. In contrast to standard TTA, continual TTA considers not only a single domain shift, but a sequence of shifts. Gradual TTA further exploits the property that some shifts evolve gradually over time. Since in both settings long test sequences are present, error accumulation needs to be addressed for methods relying on self-training. In this work, we propose and show that in the setting of TTA, the symmetric cross-entropy is better suited as a consistency loss for mean teachers compared to the commonly used cross-entropy. This is justified by our analysis with respect to the (symmetric) cross-entropy's gradient properties. To pull the test feature space closer to the source domain, where the pre-trained model is well posed, contrastive learning is leveraged. Since applications differ in their requirements, we address several settings, including having source data available and the more challenging source-free setting. We demonstrate the effectiveness of our proposed method "robust mean teacher" (RMT) on the continual and gradual corruption benchmarks CIFAR10C, CIFAR100C, and Imagenet-C. We further consider ImageNet-R and propose a new continual DomainNet-126 benchmark. State-of-the-art results are achieved on all benchmarks.
****
Understanding Imbalanced Semantic Segmentation Through Neural Collapse
Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19550-19560

A recent study has shown a phenomenon called neural collapse in that the within-class means of features and the classifier weight vectors converge to the vertices of a simplex equiangular tight frame at the terminal phase of training for classification. In this paper, we explore the corresponding structures of the last-layer feature centers and classifiers in semantic segmentation. Based on our empirical and theoretical analysis, we point out that semantic segmentation naturally brings contextual correlation and imbalanced distribution among classes, which breaks the equiangular and maximally separated structure of neural collapse for both feature centers and classifiers. However, such a symmetric structure is beneficial to discrimination for the minor classes. To preserve these advantages, we introduce a regularizer on feature centers to encourage the network to learn features closer to the appealing structure in imbalanced semantic segmentation. Experimental results show that our method can bring significant improvements on both 2D and 3D semantic segmentation benchmarks. Moreover, our method ranks first and sets a new record (+6.8% mIoU) on the ScanNet200 test leaderboard.
****
MOVES: Manipulated Objects in Video Enable Segmentation
Richard E. L. Higgins, David F. Fouhey; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6334-6343
We present a method that uses manipulation to learn to understand the objects people hold and as well as hand-object contact. We train a system that takes a single RGB image and produces a pixel-embedding that can be used to answer grouping questions (do these two pixels go together) as well as hand-association questions (is this hand holding that pixel). Rather painstakingly annotate segmentation masks, we observe people in realistic video data. We show that pairing epipolar geometry with modern optical flow produces simple and effective pseudo-labels for grouping. Given people segmentations, we can further associate pixels with hands to understand contact. Our system achieves competitive results on hand and hand-held object tasks.
****
Generating Holistic 3D Human Motion From Speech
Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 469-480
This work addresses the problem of generating 3D holistic body motions from human speech. Given a speech recording, we synthesize sequences of 3D body poses, hand gestures, and facial expressions that are realistic and diverse. To achieve this, we first build a high-quality dataset of 3D holistic body meshes with synchronous speech. We then define a novel speech-to-motion generation framework in which the face, body, and hands are modeled separately. The separated modeling stems from the fact that face articulation strongly correlates with human speech, while body poses and hand gestures are less correlated. Specifically, we employ an autoencoder for face motions, and a compositional vector-quantized variational autoencoder (VQ-VAE) for the body and hand motions. The compositional VQ-VAE is key to generating diverse results. Additionally, we propose a cross conditional autoregressive model that generates body poses and hand gestures, leading to coherent and realistic motions. Extensive experiments and user studies demonstrate that our proposed approach achieves state-of-the-art performance both qualitatively and quantitatively. Our novel dataset and code will be released for research purposes.
****
NeuDA: Neural Deformable Anchor for High-Fidelity Implicit Surface Reconstruction
Bowen Cai, Jinchi Huang, Rongfei Jia, Chengfei Lv, Huan Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8476-8485
This paper studies implicit surface reconstruction leveraging differentiable ray casting. Previous works such as IDR and NeuS overlook the spatial context in 3D space when predicting and rendering the surface, thereby may fail to capture sharp local topologies such as small holes and structures. To mitigate the limitat

ion, we propose a flexible neural implicit representation leveraging hierarchical voxel grids, namely Neural Deformable Anchor (NeuDA), for high-fidelity surface reconstruction. NeuDA maintains the hierarchical anchor grids where each vertex stores a 3d position (or anchor) instead of the direct embedding (or feature). We optimize the anchor grids such that different local geometry structures can be adaptively encoded. Besides, we dig into the frequency encoding strategies and introduce a simple hierarchical positional encoding method for the hierarchical anchor structure to flexibly exploited the properties of high-frequency and low-frequency geometry and appearance. Experiments on both the DTU and BlendedMVS datasets demonstrate that NeuDA can produce promising mesh surfaces.
****
HOICLIP: Efficient Knowledge Transfer for HOI Detection With Vision-Language Models
Shan Ning, Longtian Qiu, Yongfei Liu, Xuming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23507-23517
Human-Object Interaction (HOI) detection aims to localize human-object pairs and recognize their interactions. Recently, Contrastive Language-Image Pre-training (CLIP) has shown great potential in providing interaction prior for HOI detectors via knowledge distillation. However, such approaches often rely on large-scale training data and suffer from inferior performance under few/zero-shot scenarios. In this paper, we propose a novel HOI detection framework that efficiently extracts prior knowledge from CLIP and achieves better generalization. In detail, we first introduce a novel interaction decoder to extract informative regions in the visual feature map of CLIP via a cross-attention mechanism, which is then fused with the detection backbone by a knowledge integration block for more accurate human-object pair detection. In addition, prior knowledge in CLIP text encoder is leveraged to generate a classifier by embedding HOI descriptions. To distinguish fine-grained interactions, we build a verb classifier from training data via visual semantic arithmetic and a lightweight verb representation adapter. Furthermore, we propose a training-free enhancement to exploit global HOI predictions from CLIP. Extensive experiments demonstrate that our method outperforms the state of the art by a large margin on various settings, e.g. +4.04 mAP on HICO-Det. The source code is available in https://github.com/Artanic30/HOICLIP.
****
ShadowNeuS: Neural SDF Reconstruction by Shadow Ray Supervision
Jingwang Ling, Zhibo Wang, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 175-185
By supervising camera rays between a scene and multi-view image planes, NeRF reconstructs a neural scene representation for the task of novel view synthesis. On the other hand, shadow rays between the light source and the scene have yet to be considered. Therefore, we propose a novel shadow ray supervision scheme that optimizes both the samples along the ray and the ray location. By supervising shadow rays, we successfully reconstruct a neural SDF of the scene from single-view images under multiple lighting conditions. Given single-view binary shadows, we train a neural network to reconstruct a complete scene not limited by the camera's line of sight. By further modeling the correlation between the image colors and the shadow rays, our technique can also be effectively extended to RGB inputs. We compare our method with previous works on challenging tasks of shape reconstruction from single-view binary shadow or RGB images and observe significant improvements. The code and data are available at https://github.com/gerwang/ShadowNeuS.
****
Generalized UAV Object Detection via Frequency Domain Disentanglement
Kunyu Wang, Xueyang Fu, Yukun Huang, Chengzhi Cao, Gege Shi, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1064-1073
When deploying the Unmanned Aerial Vehicles object detection (UAV-OD) network to complex and unseen real-world scenarios, the generalization ability is usually reduced due to the domain shift. To address this issue, this paper proposes a novel frequency domain disentanglement method to improve the UAV-OD generalization

. Specifically, we first verified that the spectrum of different bands in the im
age has different effects to the UAV-OD generalization. Based on this conclusion
, we design two learnable filters to extract domain-invariant spectrum and domai
n-specific spectrum, respectively. The former can be used to train the UAV-OD ne
twork and improve its capacity for generalization. In addition, we design a new
instance-level contrastive loss to guide the network training. This loss enables
 the network to concentrate on extracting domain-invariant spectrum and domain-s
pecific spectrum, so as to achieve better disentangling results. Experimental re
sults on three unseen target domains demonstrate that our method has better gene
ralization ability than both the baseline method and state-of-the-art methods.
****
Boosting Weakly-Supervised Temporal Action Localization With Text Information
Guozhang Li, De Cheng, Xinpeng Ding, Nannan Wang, Xiaoyu Wang, Xinbo Gao; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 10648-10657
Due to the lack of temporal annotation, current Weakly-supervised Temporal Actio
n Localization (WTAL) methods are generally stuck into over-complete or incomple
te localization. In this paper, we aim to leverage the text information to boost
 WTAL from two aspects, i.e., (a) the discriminative objective to enlarge the in
ter-class difference, thus reducing the over-complete; (b) the generative object
ive to enhance the intra-class integrity, thus finding more complete temporal bo
undaries. For the discriminative objective, we propose a Text-Segment Mining (TS
M) mechanism, which constructs a text description based on the action class labe
l, and regards the text as the query to mine all class-related segments. Without
 the temporal annotation of actions, TSM compares the text query with the entire
 videos across the dataset to mine the best matching segments while ignoring irr
elevant ones. Due to the shared sub-actions in different categories of videos, m
erely applying TSM is too strict to neglect the semantic-related segments, which
 results in incomplete localization. We further introduce a generative objective
 named Video-text Language Completion (VLC), which focuses on all semantic-relat
ed segments from videos to complete the text sentence. We achieve the state-of-t
he-art performance on THUMOS14 and ActivityNet1.3. Surprisingly, we also find ou
r proposed method can be seamlessly applied to existing methods, and improve the
ir performances with a clear margin. The code is available at https://github.com
/lgzlIlIlI/Boosting-WTAL.
****
DINER: Disorder-Invariant Implicit Neural Representation
Shaowen Xie, Hao Zhu, Zhen Liu, Qi Zhang, You Zhou, Xun Cao, Zhan Ma; Proceeding
s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
2023, pp. 6143-6152
Implicit neural representation (INR) characterizes the attributes of a signal as
 a function of corresponding coordinates which emerges as a sharp weapon for sol
ving inverse problems. However, the capacity of INR is limited by the spectral b
ias in the network training. In this paper, we find that such a frequency-relate
d problem could be largely solved by re-arranging the coordinates of the input s
ignal, for which we propose the disorder-invariant implicit neural representatio
n (DINER) by augmenting a hash-table to a traditional INR backbone. Given discre
te signals sharing the same histogram of attributes and different arrangement or
ders, the hash-table could project the coordinates into the same distribution fo
r which the mapped signal can be better modeled using the subsequent INR network
, leading to significantly alleviated spectral bias. Experiments not only reveal
 the generalization of the DINER for different INR backbones (MLP vs. SIREN) and
 various tasks (image/video representation, phase retrieval, and refractive inde
x recovery) but also show the superiority over the state-of-the-art algorithms b
oth in quality and speed.
****
A Light Touch Approach to Teaching Transformers Multi-View Geometry
Yash Bhalgat, João F. Henriques, Andrew Zisserman; Proceedings of the IEEE/CVF C
onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4958-4969
Transformers are powerful visual learners, in large part due to their conspicuou

s lack of manually-specified priors. This flexibility can be problematic in tasks that involve multiple-view geometry, due to the near-infinite possible variations in 3D shapes and viewpoints (requiring flexibility), and the precise nature of projective geometry (obeying rigid laws). To resolve this conundrum, we propose a "light touch" approach, guiding visual Transformers to learn multiple-view geometry but allowing them to break free when needed. We achieve this by using epipolar lines to guide the Transformer's cross-attention maps, penalizing attention values outside the epipolar lines and encouraging higher attention along these lines since they contain geometrically plausible matches. Unlike previous methods, our proposal does not require any camera pose information at test-time. We focus on pose-invariant object instance retrieval, where standard Transformer networks struggle, due to the large differences in viewpoint between query and retrieved images. Experimentally, our method outperforms state-of-the-art approaches at object retrieval, without needing pose information at test-time.
****
Trade-Off Between Robustness and Accuracy of Vision Transformers
Yanxi Li, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7558-7568
Although deep neural networks (DNNs) have shown great successes in computer vision tasks, they are vulnerable to perturbations on inputs, and there exists a trade-off between the natural accuracy and robustness to such perturbations, which is mainly caused by the existence of robust non-predictive features and non-robust predictive features. Recent empirical analyses find Vision Transformers (ViTs) are inherently robust to various kinds of perturbations, but the aforementioned trade-off still exists for them. In this work, we propose Trade-off between Robustness and Accuracy of Vision Transformers (TORA-ViTs), which aims to efficiently transfer ViT models pretrained on natural tasks for both accuracy and robustness. TORA-ViTs consist of two major components, including a pair of accuracy and robustness adapters to extract predictive and robust features, respectively, and a gated fusion module to adjust the trade-off. The gated fusion module takes outputs of a pretrained ViT block as queries and outputs of our adapters as keys and values, and tokens from different adapters at different spatial locations are compared with each other to generate attention scores for a balanced mixing of predictive and robust features. Experiments on ImageNet with various robust benchmarks show that our TORA-ViTs can efficiently improve the robustness of naturally pretrained ViTs while maintaining competitive natural accuracy. Our most balanced setting (TORA-ViTs with lambda = 0.5) can maintain 83.7% accuracy on clean ImageNet and reach 54.7% and 38.0% accuracy under FGSM and PGD white-box attacks, respectively. In terms of various ImageNet variants, it can reach 39.2% and 56.3% accuracy on ImageNet-A and ImageNet-R and reach 34.4% mCE on ImageNet-C.
****
Focused and Collaborative Feedback Integration for Interactive Image Segmentation
Qiaoqiao Wei, Hui Zhang, Jun-Hai Yong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18643-18652
Interactive image segmentation aims at obtaining a segmentation mask for an image using simple user annotations. During each round of interaction, the segmentation result from the previous round serves as feedback to guide the user's annotation and provides dense prior information for the segmentation model, effectively acting as a bridge between interactions. Existing methods overlook the importance of feedback or simply concatenate it with the original input, leading to underutilization of feedback and an increase in the number of required annotations. To address this, we propose an approach called Focused and Collaborative Feedback Integration (FCFI) to fully exploit the feedback for click-based interactive image segmentation. FCFI first focuses on a local area around the new click and corrects the feedback based on the similarities of high-level features. It then alternately and collaboratively updates the feedback and deep features to integrate the feedback into the features. The efficacy and efficiency of FCFI were validated on four benchmarks, namely GrabCut, Berkeley, SBD, and DAVIS. Experimental results show that FCFI achieved new state-of-the-art performance with less com

putational overhead than previous methods. The source code is available at https://github.com/veizgyauzgyauz/FCFI.
****
Class Prototypes Based Contrastive Learning for Classifying Multi-Label and Fine-Grained Educational Videos
Rohit Gupta, Anirban Roy, Claire Christensen, Sujeong Kim, Sarah Gerard, Madeline Cincebeaux, Ajay Divakaran, Todd Grindal, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19923-19933
The recent growth in the consumption of online media by children during early childhood necessitates data-driven tools enabling educators to filter out appropriate educational content for young learners. This paper presents an approach for detecting educational content in online videos. We focus on two widely used educational content classes: literacy and math. For each class, we choose prominent codes (sub-classes) based on the Common Core Standards. For example, literacy codes include 'letter names', 'letter sounds', and math codes include 'counting', 'sorting'. We pose this as a fine-grained multilabel classification problem as videos can contain multiple types of educational content and the content classes can get visually similar (e.g., 'letter names' vs 'letter sounds'). We propose a novel class prototypes based supervised contrastive learning approach that can handle fine-grained samples associated with multiple labels. We learn a class prototype for each class and a loss function is employed to minimize the distances between a class prototype and the samples from the class. Similarly, distances between a class prototype and the samples from other classes are maximized. As the alignment between visual and audio cues are crucial for effective comprehension, we consider a multimodal transformer network to capture the interaction between visual and audio cues in videos while learning the embedding for videos. For evaluation, we present a dataset, APPROVE, employing educational videos from YouTube labeled with fine-grained education classes by education researchers. APPROVE consists of 193 hours of expert-annotated videos with 19 classes. The proposed approach outperforms strong baselines on APPROVE and other benchmarks such as Youtube-8M, and COIN. The dataset is available at https://nusci.csl.sri.com/project/APPROVE.
****
Deep Graph-Based Spatial Consistency for Robust Non-Rigid Point Cloud Registration
Zheng Qin, Hao Yu, Changjian Wang, Yuxing Peng, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5394-5403
We study the problem of outlier correspondence pruning for non-rigid point cloud registration. In rigid registration, spatial consistency has been a commonly used criterion to discriminate outliers from inliers. It measures the compatibility of two correspondences by the discrepancy between the respective distances in two point clouds. However, spatial consistency no longer holds in non-rigid cases and outlier rejection for non-rigid registration has not been well studied. In this work, we propose Graph-based Spatial Consistency Network (GraphSCNet) to filter outliers for non-rigid registration. Our method is based on the fact that non-rigid deformations are usually locally rigid, or local shape preserving. We first design a local spatial consistency measure over the deformation graph of the point cloud, which evaluates the spatial compatibility only between the correspondences in the vicinity of a graph node. An attention-based non-rigid correspondence embedding module is then devised to learn a robust representation of non-rigid correspondences from local spatial consistency. Despite its simplicity, GraphSCNet effectively improves the quality of the putative correspondences and attains state-of-the-art performance on three challenging benchmarks. Our code and models are available at https://github.com/qinzheng93/GraphSCNet.
****
Source-Free Adaptive Gaze Estimation by Uncertainty Reduction
Xin Cai, Jiabei Zeng, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22035-22045

Gaze estimation across domains has been explored recently because the training d
ata are usually collected under controlled conditions while the trained gaze est
imators are used in real and diverse environments. However, due to privacy and e
fficiency concerns, simultaneous access to annotated source data and to-be-predi
cted target data can be challenging. In light of this, we present an unsupervise
d source-free domain adaptation approach for gaze estimation, which adapts a sou
rce-trained gaze estimator to unlabeled target domains without source data. We p
ropose the Uncertainty Reduction Gaze Adaptation (UnReGA) framework, which achie
ves adaptation by reducing both sample and model uncertainty. Sample uncertainty
 is mitigated by enhancing image quality and making them gaze-estimation-friendl
y, whereas model uncertainty is reduced by minimizing prediction variance on the
 same inputs. Extensive experiments are conducted on six cross-domain tasks, dem
onstrating the effectiveness of UnReGA and its components. Results show that UnR
eGA outperforms other state-of-the-art cross-domain gaze estimation methods unde
r both protocols, with and without source data
****
Slide-Transformer: Hierarchical Vision Transformer With Local Self-Attention
Xuran Pan, Tianzhu Ye, Zhuofan Xia, Shiji Song, Gao Huang; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2
082-2091
Self-attention mechanism has been a key factor in the recent progress of Vision
Transformer (ViT), which enables adaptive feature extraction from global context
s. However, existing self-attention methods either adopt sparse global attention
 or window attention to reduce the computation complexity, which may compromise
the local feature learning or subject to some handcrafted designs. In contrast,
local attention, which restricts the receptive field of each query to its own ne
ighboring pixels, enjoys the benefits of both convolution and self-attention, na
mely local inductive bias and dynamic feature selection. Nevertheless, current l
ocal attention modules either use inefficient Im2Col function or rely on specifi
c CUDA kernels that are hard to generalize to devices without CUDA support. In t
his paper, we propose a novel local attention module, Slide Attention, which lev
erages common convolution operations to achieve high efficiency, flexibility and
 generalizability. Specifically, we first re-interpret the column-based Im2Col f
unction from a new row-based perspective and use Depthwise Convolution as an eff
icient substitution. On this basis, we propose a deformed shifting module based
on the re-parameterization technique, which further relaxes the fixed key/value
positions to deformed features in the local region. In this way, our module real
izes the local attention paradigm in both efficient and flexible manner. Extensi
ve experiments show that our slide attention module is applicable to a variety o
f advanced Vision Transformer models and compatible with various hardware device
s, and achieves consistently improved performances on comprehensive benchmarks.
****
NeRF-Supervised Deep Stereo
Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, Matteo Poggi; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 855-866
We introduce a novel framework for training deep stereo networks effortlessly an
d without any ground-truth. By leveraging state-of-the-art neural rendering solu
tions, we generate stereo training data from image sequences collected with a si
ngle handheld camera. On top of them, a NeRF-supervised training procedure is ca
rried out, from which we exploit rendered stereo triplets to compensate for occl
usions and depth maps as proxy labels. This results in stereo networks capable o
f predicting sharp and detailed disparity maps. Experimental results show that m
odels trained under this regime yield a 30-40% improvement over existing self-su
pervised methods on the challenging Middlebury dataset, filling the gap to super
vised models and, most times, outperforming them at zero-shot generalization.
****
Decoupled Multimodal Distilling for Emotion Recognition
Yong Li, Yuanzhi Wang, Zhen Cui; Proceedings of the IEEE/CVF Conference on Compu
ter Vision and Pattern Recognition (CVPR), 2023, pp. 6631-6640

Human multimodal emotion recognition (MER) aims to perceive human emotions via language, visual and acoustic modalities. Despite the impressive performance of previous MER approaches, the inherent multimodal heterogeneities still haunt and the contribution of different modalities varies significantly. In this work, we mitigate this issue by proposing a decoupled multimodal distillation (DMD) approach that facilitates flexible and adaptive crossmodal knowledge distillation, aiming to enhance the discriminative features of each modality. Specially, the representation of each modality is decoupled into two parts, i.e., modality-irrelevant/-exclusive spaces, in a self-regression manner. DMD utilizes a graph distillation unit (GD-Unit) for each decoupled part so that each GD can be performed in a more specialized and effective manner. A GD-Unit consists of a dynamic graph where each vertice represents a modality and each edge indicates a dynamic knowledge distillation. Such GD paradigm provides a flexible knowledge transfer manner where the distillation weights can be automatically learned, thus enabling diverse crossmodal knowledge transfer patterns. Experimental results show DMD consistently obtains superior performance than state-of-the-art MER methods. Visualization results show the graph edges in DMD exhibit meaningful distributional patterns w.r.t. the modality-irrelevant/-exclusive feature spaces. Codes are released at https://github.com/mdswyz/DMD.
****
SuperDisco: Super-Class Discovery Improves Visual Recognition for the Long-Tail
Yingjun Du, Jiayi Shen, Xiantong Zhen, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19944-19954
Modern image classifiers perform well on populated classes while degrading considerably on tail classes with only a few instances. Humans, by contrast, effortlessly handle the long-tailed recognition challenge, since they can learn the tail representation based on different levels of semantic abstraction, making the learned tail features more discriminative. This phenomenon motivated us to propose SuperDisco, an algorithm that discovers super-class representations for long-tailed recognition using a graph model. We learn to construct the super-class graph to guide the representation learning to deal with long-tailed distributions. Through message passing on the super-class graph, image representations are rectified and refined by attending to the most relevant entities based on the semantic similarity among their super-classes. Moreover, we propose to meta-learn the super-class graph under the supervision of a prototype graph constructed from a small amount of imbalanced data. By doing so, we obtain a more robust super-class graph that further improves the long-tailed recognition performance. The consistent state-of-the-art experiments on the long-tailed CIFAR-100, ImageNet, Places, and iNaturalist demonstrate the benefit of the discovered super-class graph for dealing with long-tailed distributions.
****
DualRefine: Self-Supervised Depth and Pose Estimation Through Iterative Epipolar Sampling and Refinement Toward Equilibrium
Antyanta Bangunharcana, Ahmed Magd, Kyung-Soo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 726-738
Self-supervised multi-frame depth estimation achieves high accuracy by computing matching costs of pixel correspondences between adjacent frames, injecting geometric information into the network. These pixel-correspondence candidates are computed based on the relative pose estimates between the frames. Accurate pose predictions are essential for precise matching cost computation as they influence the epipolar geometry. Furthermore, improved depth estimates can, in turn, be used to align pose estimates. Inspired by traditional structure-from-motion (SfM) principles, we propose the DualRefine model, which tightly couples depth and pose estimation through a feedback loop. Our novel update pipeline uses a deep equilibrium model framework to iteratively refine depth estimates and a hidden state of feature maps by computing local matching costs based on epipolar geometry. Importantly, we used the refined depth estimates and feature maps to compute pose updates at each step. This update in the pose estimates slowly alters the epipolar geometry during the refinement process. Experimental results on the KITTI da

taset demonstrate competitive depth prediction and odometry prediction performance surpassing published self-supervised baselines. The code is available at https://github.com/antabangun/DualRefine.
****
Improving Generalization of Meta-Learning With Inverted Regularization at Inner-Level
Lianzhe Wang, Shiji Zhou, Shanghang Zhang, Xu Chu, Heng Chang, Wenwu Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7826-7835
Despite the broad interest in meta-learning, the generalization problem remains one of the significant challenges in this field. Existing works focus on meta-generalization to unseen tasks at the meta-level by regularizing the meta-loss, while ignoring that adapted models may not generalize to the task domains at the adaptation level. In this paper, we propose a new regularization mechanism for meta-learning -- Minimax-Meta Regularization, which employs inverted regularization at the inner loop and ordinary regularization at the outer loop during training. In particular, the inner inverted regularization makes the adapted model more difficult to generalize to task domains; thus, optimizing the outer-loop loss forces the meta-model to learn meta-knowledge with better generalization. Theoretically, we prove that inverted regularization improves the meta-testing performance by reducing generalization errors. We conduct extensive experiments on the representative scenarios, and the results show that our method consistently improves the performance of meta-learning algorithms.
****
SmallCap: Lightweight Image Captioning Prompted With Retrieval Augmentation
Rita Ramos, Bruno Martins, Desmond Elliott, Yova Kementchedjhieva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2840-2849
Recent advances in image captioning have focused on scaling the data and model size, substantially increasing the cost of pre-training and finetuning. As an alternative to large models, we present SmallCap, which generates a caption conditioned on an input image and related captions retrieved from a datastore. Our model is lightweight and fast to train as the only learned parameters are in newly introduced cross-attention layers between a pre-trained CLIP encoder and GPT-2 decoder. SmallCap can transfer to new domains without additional finetuning and can exploit large-scale data in a training-free fashion since the contents of the datastore can be readily replaced. Our experiments show that SmallCap, trained only on COCO, has competitive performance on this benchmark, and also transfers to other domains without retraining, solely through retrieval from target-domain data. Further improvement is achieved through the training-free exploitation of diverse human-labeled and web data, which proves effective for a range of domains, including the nocaps benchmark, designed to test generalization to unseen visual concepts.
****
Unifying Layout Generation With a Decoupled Diffusion Model
Mude Hui, Zhizheng Zhang, Xiaoyi Zhang, Wenxuan Xie, Yuwang Wang, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1942-1951
Layout generation aims to synthesize realistic graphic scenes consisting of elements with different attributes including category, size, position, and between-element relation. It is a crucial task for reducing the burden on heavy-duty graphic design works for formatted scenes, e.g., publications, documents, and user interfaces (UIs). Diverse application scenarios impose a big challenge in unifying various layout generation subtasks, including conditional and unconditional generation. In this paper, we propose a Layout Diffusion Generative Model (LDGM) to achieve such unification with a single decoupled diffusion model. LDGM views a layout of arbitrary missing or coarse element attributes as an intermediate diffusion status from a completed layout. Since different attributes have their individual semantics and characteristics, we propose to decouple the diffusion processes for them to improve the diversity of training samples and learn the revers

e process jointly to exploit global-scope contexts for facilitating generation. As a result, our LDGM can generate layouts either from scratch or conditional on arbitrary available attributes. Extensive qualitative and quantitative experiments demonstrate our proposed LDGM outperforms existing layout generation models in both functionality and performance.
****
Im2Hands: Learning Attentive Implicit Representation of Interacting Two-Hand Shapes
Jihyun Lee, Minhyuk Sung, Honggyu Choi, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21169-21178
We present Implicit Two Hands (Im2Hands), the first neural implicit representation of two interacting hands. Unlike existing methods on two-hand reconstruction that rely on a parametric hand model and/or low-resolution meshes, Im2Hands can produce fine-grained geometry of two hands with high hand-to-hand and hand-to-image coherency. To handle the shape complexity and interaction context between two hands, Im2Hands models the occupancy volume of two hands -- conditioned on an RGB image and coarse 3D keypoints -- by two novel attention-based modules responsible for (1) initial occupancy estimation and (2) context-aware occupancy refinement, respectively. Im2Hands first learns per-hand neural articulated occupancy in the canonical space designed for each hand using query-image attention. It then refines the initial two-hand occupancy in the posed space to enhance the coherency between the two hand shapes using query-anchor attention. In addition, we introduce an optional keypoint refinement module to enable robust two-hand shape estimation from predicted hand keypoints in a single-image reconstruction scenario. We experimentally demonstrate the effectiveness of Im2Hands on two-hand reconstruction in comparison to related methods, where ours achieves state-of-the-art results. Our code is publicly available at https://github.com/jyunlee/Im2Hands.
****
Long-Term Visual Localization With Mobile Sensors
Shen Yan, Yu Liu, Long Wang, Zehong Shen, Zhen Peng, Haomin Liu, Maojun Zhang, Guofeng Zhang, Xiaowei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17245-17255
Despite the remarkable advances in image matching and pose estimation, image-based localization of a camera in a temporally-varying outdoor environment is still a challenging problem due to huge appearance disparity between query and reference images caused by illumination, seasonal and structural changes. In this work, we propose to leverage additional sensors on a mobile phone, mainly GPS, compass, and gravity sensor, to solve this challenging problem. We show that these mobile sensors provide decent initial poses and effective constraints to reduce the searching space in image matching and final pose estimation. With the initial pose, we are also able to devise a direct 2D-3D matching network to efficiently establish 2D-3D correspondences instead of tedious 2D-2D matching in existing systems. As no public dataset exists for the studied problem, we collect a new dataset that provides a variety of mobile sensor data and significant scene appearance variations, and develop a system to acquire ground-truth poses for query images. We benchmark our method as well as several state-of-the-art baselines and demonstrate the effectiveness of the proposed approach. Our code and dataset are available on the project page: https://zju3dv.github.io/sensloc/.
****
Data-Efficient Large Scale Place Recognition With Graded Similarity Supervision
María Leyva-Vallina, Nicola Strisciuglio, Nicolai Petkov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23487-23496
Visual place recognition (VPR) is a fundamental task of computer vision for visual localization. Existing methods are trained using image pairs that either depict the same place or not. Such a binary indication does not consider continuous relations of similarity between images of the same place taken from different positions, determined by the continuous nature of camera pose. The binary similari

ty induces a noisy supervision signal into the training of VPR methods, which stall in local minima and require expensive hard mining algorithms to guarantee convergence. Motivated by the fact that two images of the same place only partially share visual cues due to camera pose differences, we deploy an automatic re-annotation strategy to re-label VPR datasets. We compute graded similarity labels for image pairs based on available localization metadata. Furthermore, we propose a new Generalized Contrastive Loss (GCL) that uses graded similarity labels for training contrastive networks. We demonstrate that the use of the new labels and GCL allow to dispense from hard-pair mining, and to train image descriptors that perform better in VPR by nearest neighbor search, obtaining superior or comparable results than methods that require expensive hard-pair mining and re-ranking techniques.
****
Dynamic Neural Network for Multi-Task Learning Searching Across Diverse Network Topologies
Wonhyeok Choi, Sunghoon Im; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3779-3788
In this paper, we present a new MTL framework that searches for structures optimized for multiple tasks with diverse graph topologies and shares features among tasks. We design a restricted DAG-based central network with read-in/read-out layers to build topologically diverse task-adaptive structures while limiting search space and time. We search for a single optimized network that serves as multiple task adaptive sub-networks using our three-stage training process. To make the network compact and discretized, we propose a flow-based reduction algorithm and a squeeze loss used in the training process. We evaluate our optimized network on various public MTL datasets and show ours achieves state-of-the-art performance. An extensive ablation study experimentally validates the effectiveness of the sub-module and schemes in our framework.
****
Relightable Neural Human Assets From Multi-View Gradient Illuminations
Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, Jingyi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4315-4327
Human modeling and relighting are two fundamental problems in computer vision and graphics, where high-quality datasets can largely facilitate related research. However, most existing human datasets only provide multi-view human images captured under the same illumination. Although valuable for modeling tasks, they are not readily used in relighting problems. To promote research in both fields, in this paper, we present UltraStage, a new 3D human dataset that contains more than 2,000 high-quality human assets captured under both multi-view and multi-illumination settings. Specifically, for each example, we provide 32 surrounding views illuminated with one white light and two gradient illuminations. In addition to regular multi-view images, gradient illuminations help recover detailed surface normal and spatially-varying material maps, enabling various relighting applications. Inspired by recent advances in neural representation, we further interpret each example into a neural human asset which allows novel view synthesis under arbitrary lighting conditions. We show our neural human assets can achieve extremely high capture performance and are capable of representing fine details such as facial wrinkles and cloth folds. We also validate UltraStage in single image relighting tasks, training neural networks with virtual relighted data from neural assets and demonstrating realistic rendering improvements over prior arts. UltraStage will be publicly available to the community to stimulate significant future developments in various human modeling and rendering tasks. The dataset is available at https://miaoing.github.io/RNHA.
****
Probing Sentiment-Oriented Pre-Training Inspired by Human Sentiment Perception Mechanism
Tinglei Feng, Jiaxuan Liu, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2850-2860
Pre-training of deep convolutional neural networks (DCNNs) plays a crucial role

in the field of visual sentiment analysis (VSA). Most proposed methods employ the off-the-shelf backbones pre-trained on large-scale object classification datasets (i.e., ImageNet). While it boosts performance for a big margin against initializing model states from random, we argue that DCNNs simply pre-trained on ImageNet may excessively focus on recognizing objects, but failed to provide high-level concepts in terms of sentiment. To address this long-term overlooked problem, we propose a sentiment-oriented pre-training method that is built upon human visual sentiment perception (VSP) mechanism. Specifically, we factorize the process of VSP into three steps, namely stimuli taking, holistic organizing, and high-level perceiving. From imitating each VSP step, a total of three models are separately pre-trained via our devised sentiment-aware tasks that contribute to excavating sentiment-discriminated representations. Moreover, along with our elaborated multi-model amalgamation strategy, the prior knowledge learned from each perception step can be effectively transferred into a single target model, yielding substantial performance gains. Finally, we verify the superiorities of our proposed method over extensive experiments, covering mainstream VSA tasks from single-label learning (SLL), multi-label learning (MLL), to label distribution learning (LDL). Experiment results demonstrate that our proposed method leads to unanimous improvements in these downstream tasks. Our code is released on https://github.com/tinglyfeng/sentiment_pretraining
****
Imitation Learning As State Matching via Differentiable Physics
Siwei Chen, Xiao Ma, Zhongwen Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7846-7855
Existing imitation learning (IL) methods such as inverse reinforcement learning (IRL) usually have a double-loop training process, alternating between learning a reward function and a policy and tend to suffer long training time and high variance. In this work, we identify the benefits of differentiable physics simulators and propose a new IL method, i.e., Imitation Learning via Differentiable Physics (ILD), which gets rid of the double-loop design and achieves significant improvements in final performance, convergence speed, and stability. The proposed ILD incorporates the differentiable physics simulator as a physics prior into its computational graph for policy learning. It unrolls the dynamics by sampling actions from a parameterized policy, simply minimizing the distance between the expert trajectory and the agent trajectory, and back-propagating the gradient into the policy via temporal physics operators. With the physics prior, ILD policies can not only be transferable to unseen environment specifications but also yield higher final performance on a variety of tasks. In addition, ILD naturally forms a single-loop structure, which significantly improves the stability and training speed. To simplify the complex optimization landscape induced by temporal physics operations, ILD dynamically selects the learning objectives for each state during optimization. In our experiments, we show that ILD outperforms state-of-the-art methods in a variety of continuous control tasks with Brax, requiring only one expert demonstration. In addition, ILD can be applied to challenging deformable object manipulation tasks and can be generalized to unseen configurations.
****
OpenMix: Exploring Outlier Samples for Misclassification Detection
Fei Zhu, Zhen Cheng, Xu-Yao Zhang, Cheng-Lin Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12074-12083
Reliable confidence estimation for deep neural classifiers is a challenging yet fundamental requirement in high-stakes applications. Unfortunately, modern deep neural networks are often overconfident for their erroneous predictions. In this work, we exploit the easily available outlier samples, i.e., unlabeled samples coming from non-target classes, for helping detect misclassification errors. Particularly, we find that the well-known Outlier Exposure, which is powerful in detecting out-of-distribution (OOD) samples from unknown classes, does not provide any gain in identifying misclassification errors. Based on these observations, we propose a novel method called OpenMix, which incorporates open-world knowledg

e by learning to reject uncertain pseudo-samples generated via outlier transformation. OpenMix significantly improves confidence reliability under various scenarios, establishing a strong and unified framework for detecting both misclassified samples from known classes and OOD samples from unknown classes.
****
Multivariate, Multi-Frequency and Multimodal: Rethinking Graph Neural Networks for Emotion Recognition in Conversation
Feiyu Chen, Jie Shao, Shuyuan Zhu, Heng Tao Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10761-10770
Complex relationships of high arity across modality and context dimensions is a critical challenge in the Emotion Recognition in Conversation (ERC) task. Yet, previous works tend to encode multimodal and contextual relationships in a loosely-coupled manner, which may harm relationship modelling. Recently, Graph Neural Networks (GNN) which show advantages in capturing data relations, offer a new solution for ERC. However, existing GNN-based ERC models fail to address some general limits of GNNs, including assuming pairwise formulation and erasing high-frequency signals, which may be trivial for many applications but crucial for the ERC task. In this paper, we propose a GNN-based model that explores multivariate relationships and captures the varying importance of emotion discrepancy and commonality by valuing multi-frequency signals. We empower GNNs to better capture the inherent relationships among utterances and deliver more sufficient multimodal and contextual modelling. Experimental results show that our proposed method outperforms previous state-of-the-art works on two popular multimodal ERC datasets.
****
Weakly Supervised Class-Agnostic Motion Prediction for Autonomous Driving
Ruibo Li, Hanyu Shi, Ziang Fu, Zhe Wang, Guosheng Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17599-17608
Understanding the motion behavior of dynamic environments is vital for autonomous driving, leading to increasing attention in class-agnostic motion prediction in LiDAR point clouds. Outdoor scenes can often be decomposed into mobile foregrounds and static backgrounds, which enables us to associate motion understanding with scene parsing. Based on this observation, we study a novel weakly supervised motion prediction paradigm, where fully or partially (1%, 0.1%) annotated foreground/background binary masks rather than expensive motion annotations are used for supervision. To this end, we propose a two-stage weakly supervised approach, where the segmentation model trained with the incomplete binary masks in Stage1 will facilitate the self-supervised learning of the motion prediction network in Stage2 by estimating possible moving foregrounds in advance. Furthermore, for robust self-supervised motion learning, we design a Consistency-aware Chamfer Distance loss by exploiting multi-frame information and explicitly suppressing potential outliers. Comprehensive experiments show that, with fully or partially binary masks as supervision, our weakly supervised models surpass the self-supervised models by a large margin and perform on par with some supervised ones. This further demonstrates that our approach achieves a good compromise between annotation effort and performance.
****
TOPLight: Lightweight Neural Networks With Task-Oriented Pretraining for Visible-Infrared Recognition
Hao Yu, Xu Cheng, Wei Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3541-3550
Visible-infrared recognition (VI recognition) is a challenging task due to the enormous visual difference across heterogeneous images. Most existing works achieve promising results by transfer learning, such as pretraining on the ImageNet, based on advanced neural architectures like ResNet and ViT. However, such methods ignore the negative influence of the pretrained colour prior knowledge, as well as their heavy computational burden makes them hard to deploy in actual scenarios with limited resources. In this paper, we propose a novel task-oriented pret

rained lightweight neural network (TOPLight) for VI recognition. Specifically, t
he TOPLight method simulates the domain conflict and sample variations with the
proposed fake domain loss in the pretraining stage, which guides the network to
learn how to handle those difficulties, such that a more general modality-shared
 feature representation is learned for the heterogeneous images. Moreover, an ef
fective fine-grained dependency reconstruction module (FDR) is developed to disc
over substantial pattern dependencies shared in two modalities. Extensive experi
ments on VI person re-identification and VI face recognition datasets demonstrat
e the superiority of the proposed TOPLight, which significantly outperforms the
current state of the arts while demanding fewer computational resources.
****
DeFeeNet: Consecutive 3D Human Motion Prediction With Deviation Feedback
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)
, 2023, pp. 5527-5536
Let us rethink the real-world scenarios that require human motion prediction tec
hniques, such as human-robot collaboration. Current works simplify the task of p
redicting human motions into a one-off process of forecasting a short future seq
uence (usually no longer than 1 second) based on a historical observed one. Howe
ver, such simplification may fail to meet practical needs due to the neglect of
the fact that motion prediction in real applications is not an isolated "observe
 then predict" unit, but a consecutive process composed of many rounds of such u
nit, semi-overlapped along the entire sequence. As time goes on, the predicted p
art of previous round has its corresponding ground truth observable in the new r
ound, but their deviation in-between is neither exploited nor able to be capture
d by existing isolated learning fashion. In this paper, we propose DeFeeNet, a s
imple yet effective network that can be added on existing one-off prediction mod
els to realize deviation perception and feedback when applied to consecutive mot
ion prediction task. At each prediction round, the deviation generated by previo
us unit is first encoded by our DeFeeNet, and then incorporated into the existin
g predictor to enable a deviation-aware prediction manner, which, for the first
time, allows for information transmit across adjacent prediction units. We desig
n two versions of DeFeeNet as MLP-based and GRU-based, respectively. On Human3.6
M and more complicated BABEL, experimental results indicate that our proposed ne
twork improves consecutive human motion prediction performance regardless of the
 basic model.
****
Where We Are and What We're Looking At: Query Based Worldwide Image Geo-Localiza
tion Using Hierarchies and Scenes
Brandon Clark, Alec Kerrigan, Parth Parag Kulkarni, Vicente Vivanco Cepeda, Muba
rak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2023, pp. 23182-23190
Determining the exact latitude and longitude that a photo was taken is a useful
and widely applicable task, yet it remains exceptionally difficult despite the a
ccelerated progress of other computer vision tasks. Most previous approaches hav
e opted to learn single representations of query images, which are then classifi
ed at different levels of geographic granularity. These approaches fail to explo
it the different visual cues that give context to different hierarchies, such as
 the country, state, and city level. To this end, we introduce an end-to-end tra
nsformer-based architecture that exploits the relationship between different geo
graphic levels (which we refer to as hierarchies) and the corresponding visual s
cene information in an image through hierarchical cross-attention. We achieve th
is by learning a query for each geographic hierarchy and scene type. Furthermore
, we learn a separate representation for different environmental scenes, as diff
erent scenes in the same location are often defined by completely different visu
al features. We achieve state of the art accuracy on 4 standard geo-localization
 datasets : Im2GPS, Im2GPS3k, YFCC4k, and YFCC26k, as well as qualitatively demo
nstrate how our method learns different representations for different visual hie
rarchies and scenes, which has not been demonstrated in the previous methods. Ab
ove previous testing datasets mostly consist of iconic landmarks or images taken

from social media, which makes the dataset a simple memory task, or makes it biased towards certain places. To address this issue we introduce a much harder testing dataset, Google-World-Streets-15k, comprised of images taken from Google Streetview covering the whole planet and present state of the art results. Our code can be found at https://github.com/AHKerrigan/GeoGuessNet.
****

Bridging Precision and Confidence: A Train-Time Loss for Calibrating Object Detection

Muhammad Akhtar Munir, Muhammad Haris Khan, Salman Khan, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11474-11483

Deep neural networks (DNNs) have enabled astounding progress in several vision-based problems. Despite showing high predictive accuracy, recently, several works have revealed that they tend to provide overconfident predictions and thus are poorly calibrated. The majority of the works addressing the miscalibration of DNNs fall under the scope of classification and consider only in-domain predictions. However, there is little to no progress in studying the calibration of DNN-based object detection models, which are central to many vision-based safety-critical applications. In this paper, inspired by the train-time calibration methods, we propose a novel auxiliary loss formulation that explicitly aims to align the class confidence of bounding boxes with the accurateness of predictions (i.e. precision). Since the original formulation of our loss depends on the counts of true positives and false positives in a minibatch, we develop a differentiable proxy of our loss that can be used during training with other application-specific loss functions. We perform extensive experiments on challenging in-domain and out-domain scenarios with six benchmark datasets including MS-COCO, Cityscapes, Sim10k, and BDD100k. Our results reveal that our train-time loss surpasses strong calibration baselines in reducing calibration error for both in and out-domain scenarios. Our source code and pre-trained models are available at https://github.com/akhtarvision/bpc_calibration
****

DyLiN: Making Light Field Networks Dynamic

Heng Yu, Joel Julin, Zoltán Á. Milacski, Koichiro Niinuma, László A. Jeni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12397-12406

Light Field Networks, the re-formulations of radiance fields to oriented rays, are magnitudes faster than their coordinate network counterparts, and provide higher fidelity with respect to representing 3D structures from 2D observations. They would be well suited for generic scene representation and manipulation, but suffer from one problem: they are limited to holistic and static scenes. In this paper, we propose the Dynamic Light Field Network (DyLiN) method that can handle non-rigid deformations, including topological changes. We learn a deformation field from input rays to canonical rays, and lift them into a higher dimensional space to handle discontinuities. We further introduce CoDyLiN, which augments DyLiN with controllable attribute inputs. We train both models via knowledge distillation from pretrained dynamic radiance fields. We evaluated DyLiN using both synthetic and real world datasets that include various non-rigid deformations. DyLiN qualitatively outperformed and quantitatively matched state-of-the-art methods in terms of visual fidelity, while being 25 - 71x computationally faster. We also tested CoDyLiN on attribute annotated data and it surpassed its teacher model. Project page: https://dylin2023.github.io.
****

Critical Learning Periods for Multisensory Integration in Deep Networks

Michael Kleinman, Alessandro Achille, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24296-24305

We show that the ability of a neural network to integrate information from diverse sources hinges critically on being exposed to properly correlated signals during the early phases of training. Interfering with the learning process during this initial stage can permanently impair the development of a skill, both in art

ificial and biological systems where the phenomenon is known as a critical learn
ing period. We show that critical periods arise from the complex and unstable ea
rly transient dynamics, which are decisive of final performance of the trained s
ystem and their learned representations. This evidence challenges the view, enge
ndered by analysis of wide and shallow networks, that early learning dynamics of
 neural networks are simple, akin to those of a linear model. Indeed, we show th
at even deep linear networks exhibit critical learning periods for multi-source
integration, while shallow networks do not. To better understand how the interna
l representations change according to disturbances or sensory deficits, we intro
duce a new measure of source sensitivity, which allows us to track the inhibitio
n and integration of sources during training. Our analysis of inhibition suggest
s cross-source reconstruction as a natural auxiliary training objective, and ind
eed we show that architectures trained with cross-sensor reconstruction objectiv
es are remarkably more resilient to critical periods. Our findings suggest that
the recent success in self-supervised multi-modal training compared to previous
supervised efforts may be in part due to more robust learning dynamics and not s
olely due to better architectures and/or more data.
****
Human Guided Ground-Truth Generation for Realistic Image Super-Resolution
Du Chen, Jie Liang, Xindong Zhang, Ming Liu, Hui Zeng, Lei Zhang; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 14082-14091
How to generate the ground-truth (GT) image is a critical issue for training rea
listic image super-resolution (Real-ISR) models. Existing methods mostly take a
set of high-resolution (HR) images as GTs and apply various degradations to simu
late their low-resolution (LR) counterparts. Though great progress has been achi
eved, such an LR-HR pair generation scheme has several limitations. First, the p
erceptual quality of HR images may not be high enough, limiting the quality of R
eal-ISR outputs. Second, existing schemes do not consider much human perception
in GT generation, and the trained models tend to produce over-smoothed results o
r unpleasant artifacts. With the above considerations, we propose a human guided
 GT generation scheme. We first elaborately train multiple image enhancement mod
els to improve the perceptual quality of HR images, and enable one LR image havi
ng multiple HR counterparts. Human subjects are then involved to annotate the hi
gh quality regions among the enhanced HR images as GTs, and label the regions wi
th unpleasant artifacts as negative samples. A human guided GT image dataset wit
h both positive and negative samples is then constructed, and a loss function is
 proposed to train the Real-ISR models. Experiments show that the Real-ISR model
s trained on our dataset can produce perceptually more realistic results with le
ss artifacts. Dataset and codes can be found at https://github.com/ChrisDud0257/
HGGT.
****
GarmentTracking: Category-Level Garment Pose Tracking
Han Xue, Wenqiang Xu, Jieyi Zhang, Tutian Tang, Yutong Li, Wenxin Du, Ruolin Ye,
 Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2023, pp. 21233-21242
Garments are important to humans. A visual system that can estimate and track th
e complete garment pose can be useful for many downstream tasks and real-world a
pplications. In this work, we present a complete package to address the category
-level garment pose tracking task: (1) A recording system VR-Garment, with which
 users can manipulate virtual garment models in simulation through a VR interfac
e. (2) A large-scale dataset VR-Folding, with complex garment pose configuration
s in manipulation like flattening and folding. (3) An end-to-end online tracking
 framework GarmentTracking, which predicts complete garment pose both in canonic
al space and task space given a point cloud sequence. Extensive experiments demo
nstrate that the proposed GarmentTracking achieves great performance even when t
he garment has large non-rigid deformation. It outperforms the baseline approach
 on both speed and accuracy. We hope our proposed solution can serve as a platfo
rm for future research. Codes and datasets are available in https://garment-trac
king.robotflow.ai.

****
Mask DINO: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation
Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M. Ni, Heung-Yeung Shum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3041-3050
In this paper we present Mask DINO, a unified object detection and segmentation framework. Mask DINO extends DINO (DETR with Improved Denoising Anchor Boxes) by adding a mask prediction branch which supports all image segmentation tasks (instance, panoptic, and semantic). It makes use of the query embeddings from DINO to dot-product a high-resolution pixel embedding map to predict a set of binary masks. Some key components in DINO are extended for segmentation through a shared architecture and training process. Mask DINO is simple, efficient, scalable, and benefits from joint large-scale detection and segmentation datasets. Our experiments show that Mask DINO significantly outperforms all existing specialized segmentation methods, both on a ResNet-50 backbone and a pre-trained model with SwinL backbone. Notably, Mask DINO establishes the best results to date on instance segmentation (54.5 AP on COCO), panoptic segmentation (59.4 PQ on COCO), and semantic segmentation (60.8 mIoU on ADE20K) among models under one billion parameters. We will release the code after the blind review.
****
Align and Attend: Multimodal Summarization With Dual Contrastive Losses
Bo He, Jun Wang, Jielin Qiu, Trung Bui, Abhinav Shrivastava, Zhaowen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14867-14878
The goal of multimodal summarization is to extract the most important information from different modalities to form summaries. Unlike unimodal summarization, the multimodal summarization task explicitly leverages cross-modal information to help generate more reliable and high-quality summaries. However, existing methods fail to leverage the temporal correspondence between different modalities and ignore the intrinsic correlation between different samples. To address this issue, we introduce Align and Attend Multimodal Summarization (A2Summ), a unified multimodal transformer-based model which can effectively align and attend the multimodal input. In addition, we propose two novel contrastive losses to model both inter-sample and intra-sample correlations. Extensive experiments on two standard video summarization datasets (TVSum and SumMe) and two multimodal summarization datasets (Daily Mail and CNN) demonstrate the superiority of A2Summ, achieving state-of-the-art performances on all datasets. Moreover, we collected a large-scale multimodal summarization dataset BLiSS, which contains livestream videos and transcribed texts with annotated summaries. Our code and dataset are publicly available at https://boheumd.github.io/A2Summ/.
****
SinGRAF: Learning a 3D Generative Radiance Field for a Single Scene
Minjung Son, Jeong Joon Park, Leonidas Guibas, Gordon Wetzstein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8507-8517
Generative models have shown great promise in synthesizing photorealistic 3D objects, but they require large amounts of training data. We introduce SinGRAF, a 3D-aware generative model that is trained with a few input images of a single scene. Once trained, SinGRAF generates different realizations of this 3D scene that preserve the appearance of the input while varying scene layout. For this purpose, we build on recent progress in 3D GAN architectures and introduce a novel progressive-scale patch discrimination approach during training. With several experiments, we demonstrate that the results produced by SinGRAF outperform the closest related works in both quality and diversity by a large margin.
****
Self-Supervised AutoFlow
Hsin-Ping Huang, Charles Herrmann, Junhwa Hur, Erika Lu, Kyle Sargent, Austin Stone, Ming-Hsuan Yang, Deqing Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11412-11421

Recently, AutoFlow has shown promising results on learning a training set for optical flow, but requires ground truth labels in the target domain to compute its search metric. Observing a strong correlation between the ground truth search metric and self-supervised losses, we introduce self-supervised AutoFlow to handle real-world videos without ground truth labels. Using self-supervised loss as the search metric, our self-supervised AutoFlow performs on par with AutoFlow on Sintel and KITTI where ground truth is available, and performs better on the real-world DAVIS dataset. We further explore using self-supervised AutoFlow in the (semi-)supervised setting and obtain competitive results against the state of the art.
****
MagicNet: Semi-Supervised Multi-Organ Segmentation via Magic-Cube Partition and Recovery
Duowen Chen, Yunhao Bai, Wei Shen, Qingli Li, Lequan Yu, Yan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23869-23878
We propose a novel teacher-student model for semi-supervised multi-organ segmentation. In the teacher-student model, data augmentation is usually adopted on unlabeled data to regularize the consistent training between teacher and student. We start from a key perspective that fixed relative locations and variable sizes of different organs can provide distribution information where a multi-organ CT scan is drawn. Thus, we treat the prior anatomy as a strong tool to guide the data augmentation and reduce the mismatch between labeled and unlabeled images for semi-supervised learning. More specifically, we propose a data augmentation strategy based on partition-and-recovery N^3 cubes cross- and within- labeled and unlabeled images. Our strategy encourages unlabeled images to learn organ semantics in relative locations from the labeled images (cross-branch) and enhances the learning ability for small organs (within-branch). For within-branch, we further propose to refine the quality of pseudo labels by blending the learned representations from small cubes to incorporate local attributes. Our method is termed as MagicNet, since it treats the CT volume as a magic-cube and N^3-cube partition-and-recovery process matches with the rule of playing a magic-cube. Extensive experiments on two public CT multi-organ datasets demonstrate the effectiveness of MagicNet, and noticeably outperforms state-of-the-art semi-supervised medical image segmentation approaches, with +7% DSC improvement on MACT dataset with 10% labeled images.
****
Neuralangelo: High-Fidelity Neural Surface Reconstruction
Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H. Taylor, Mathias Unberath, Ming-Yu Liu, Chen-Hsuan Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8456-8465
Neural surface reconstruction has been shown to be powerful for recovering dense 3D surfaces via image-based neural rendering. However, current methods struggle to recover detailed structures of real-world scenes. To address the issue, we present Neuralangelo, which combines the representation power of multi-resolution 3D hash grids with neural surface rendering. Two key ingredients enable our approach: (1) numerical gradients for computing higher-order derivatives as a smoothing operation and (2) coarse-to-fine optimization on the hash grids controlling different levels of details. Even without auxiliary inputs such as depth, Neuralangelo can effectively recover dense 3D surface structures from multi-view images with fidelity significantly surpassing previous methods, enabling detailed large-scale scene reconstruction from RGB video captures.
****
Re-GAN: Data-Efficient GANs Training via Architectural Reconfiguration
Divya Saxena, Jiannong Cao, Jiahao Xu, Tarun Kulshrestha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16230-16240
Training Generative Adversarial Networks (GANs) on high-fidelity images usually requires a vast number of training images. Recent research on GAN tickets reveals that dense GANs models contain sparse sub-networks or "lottery tickets" that,

when trained separately, yield better results under limited data. However, finding GANs tickets requires an expensive process of train-prune-retrain. In this paper, we propose Re-GAN, a data-efficient GANs training that dynamically reconfigures GANs architecture during training to explore different sub-network structures in training time. Our method repeatedly prunes unimportant connections to regularize GANs network and regrows them to reduce the risk of prematurely pruning important connections. Re-GAN stabilizes the GANs models with less data and offers an alternative to the existing GANs tickets and progressive growing methods. We demonstrate that Re-GAN is a generic training methodology which achieves stability on datasets of varying sizes, domains, and resolutions (CIFAR-10, Tiny-ImageNet, and multiple few-shot generation datasets) as well as different GANs architectures (SNGAN, ProGAN, StyleGAN2 and AutoGAN). Re-GAN also improves performance when combined with the recent augmentation approaches. Moreover, Re-GAN requires fewer floating-point operations (FLOPs) and less training time by removing the unimportant connections during GANs training while maintaining comparable or even generating higher-quality samples. When compared to state-of-the-art StyleGAN2, our method outperforms without requiring any additional fine-tuning step. Code can be found at this link: https://github.com/IntellicentAI-Lab/Re-GAN
****

Dimensionality-Varying Diffusion Process
Han Zhang, Ruili Feng, Zhantao Yang, Lianghua Huang, Yu Liu, Yifei Zhang, Yujun Shen, Deli Zhao, Jingren Zhou, Fan Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14307-14316
Diffusion models, which learn to reverse a signal destruction process to generate new data, typically require the signal at each step to have the same dimension. We argue that, considering the spatial redundancy in image signals, there is no need to maintain a high dimensionality in the evolution process, especially in the early generation phase. To this end, we make a theoretical generalization of the forward diffusion process via signal decomposition. Concretely, we manage to decompose an image into multiple orthogonal components and control the attenuation of each component when perturbing the image. That way, along with the noise strength increasing, we are able to diminish those inconsequential components and thus use a lower-dimensional signal to represent the source, barely losing information. Such a reformulation allows to vary dimensions in both training and inference of diffusion models. Extensive experiments on a range of datasets suggest that our approach substantially reduces the computational cost and achieves on-par or even better synthesis performance compared to baseline methods. We also show that our strategy facilitates high-resolution image synthesis and improves FID of diffusion model trained on FFHQ at 1024x1024 resolution from 52.40 to 10.46. Code is available at https://github.com/damo-vilab/dvdp.
****

FAME-ViL: Multi-Tasking Vision-Language Model for Heterogeneous Fashion Tasks
Xiao Han, Xiatian Zhu, Licheng Yu, Li Zhang, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2669-2680
In the fashion domain, there exists a variety of vision-and-language (V+L) tasks, including cross-modal retrieval, text-guided image retrieval, multi-modal classification, and image captioning. They differ drastically in each individual input/output format and dataset size. It has been common to design a task-specific model and fine-tune it independently from a pre-trained V+L model (e.g., CLIP). This results in parameter inefficiency and inability to exploit inter-task relatedness. To address such issues, we propose a novel FAshion-focused Multi-task Efficient learning method for Vision-and-Language tasks (FAME-ViL) in this work. Compared with existing approaches, FAME-ViL applies a single model for multiple heterogeneous fashion tasks, therefore being much more parameter-efficient. It is enabled by two novel components: (1) a task-versatile architecture with cross-attention adapters and task-specific adapters integrated into a unified V+L model, and (2) a stable and effective multi-task training strategy that supports learning from heterogeneous data and prevents negative transfer. Extensive experiments on four fashion tasks show that our FAME-ViL can save 61.5% of parameters ove

r alternatives, while significantly outperforming the conventional independently trained single-task models. Code is available at https://github.com/BrandonHanx/FAME-ViL.
****
Neural Intrinsic Embedding for Non-Rigid Point Cloud Matching
Puhua Jiang, Mingze Sun, Ruqi Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21835-21845
As a primitive 3D data representation, point clouds are prevailing in 3D sensing, yet short of intrinsic structural information of the underlying objects. Such discrepancy poses great challenges in directly establishing correspondences between point clouds sampled from deformable shapes. In light of this, we propose Neural Intrinsic Embedding (NIE) to embed each vertex into a high-dimensional space in a way that respects the intrinsic structure. Based upon NIE, we further present a weakly-supervised learning framework for non-rigid point cloud registration. Unlike the prior works, we do not require expansive and sensitive off-line basis construction (e.g., eigen-decomposition of Laplacians), nor do we require ground-truth correspondence labels for supervision. We empirically show that our framework performs on par with or even better than the state-of-the-art baselines, which generally require more supervision and/or more structural geometric input.
****
Rate Gradient Approximation Attack Threats Deep Spiking Neural Networks
Tong Bu, Jianhao Ding, Zecheng Hao, Zhaofei Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7896-7906
Spiking Neural Networks (SNNs) have attracted significant attention due to their energy-efficient properties and potential application on neuromorphic hardware. State-of-the-art SNNs are typically composed of simple Leaky Integrate-and-Fire (LIF) neurons and have become comparable to ANNs in image classification tasks on large-scale datasets. However, the robustness of these deep SNNs has not yet been fully uncovered. In this paper, we first experimentally observe that layers in these SNNs mostly communicate by rate coding. Based on this rate coding property, we develop a novel rate coding SNN-specified attack method, Rate Gradient Approximation Attack (RGA). We generalize the RGA attack to SNNs composed of LIF neurons with different leaky parameters and input encoding by designing surrogate gradients. In addition, we develop the time-extended enhancement to generate more effective adversarial examples. The experiment results indicate that our proposed RGA attack is more effective than the previous attack and is less sensitive to neuron hyperparameters. We also conclude from the experiment that rate-coded SNN composed of LIF neurons is not secure, which calls for exploring training methods for SNNs composed of complex neurons and other neuronal codings. Code is available at https://github.com/putshua/SNN_attack_RGA
****
Few-Shot Geometry-Aware Keypoint Localization
Xingzhe He, Gaurav Bharaj, David Ferman, Helge Rhodin, Pablo Garrido; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21337-21348
Supervised keypoint localization methods rely on large manually labeled image datasets, where objects can deform, articulate, or occlude. However, creating such large keypoint labels is time-consuming and costly, and is often error-prone due to inconsistent labeling. Thus, we desire an approach that can learn keypoint localization with fewer yet consistently annotated images. To this end, we present a novel formulation that learns to localize semantically consistent keypoint definitions, even for occluded regions, for varying object categories. We use a few user-labeled 2D images as input examples, which are extended via self-supervision using a larger unlabeled dataset. Unlike unsupervised methods, the few-shot images act as semantic shape constraints for object localization. Furthermore, we introduce 3D geometry-aware constraints to uplift keypoints, achieving more accurate 2D localization. Our general-purpose formulation paves the way for semantically conditioned generative modeling and attains competitive or state-of-the-art accuracy on several datasets, including human faces, eyes, animals, cars, a

nd never-before-seen mouth interior (teeth) localization tasks, not attempted by
 the previous few-shot methods. Project page: https://xingzhehe.github.io/FewSho
t3DKP/
****
RenderDiffusion: Image Diffusion for 3D Reconstruction, Inpainting and Generatio
n
Titas Anciukevi■ius, Zexiang Xu, Matthew Fisher, Paul Henderson, Hakan Bilen, Ni
loy J. Mitra, Paul Guerrero; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2023, pp. 12608-12618
Diffusion models currently achieve state-of-the-art performance for both conditi
onal and unconditional image generation. However, so far, image diffusion models
 do not support tasks required for 3D understanding, such as view-consistent 3D
generation or single-view object reconstruction. In this paper, we present Rende
rDiffusion, the first diffusion model for 3D generation and inference, trained u
sing only monocular 2D supervision. Central to our method is a novel image denoi
sing architecture that generates and renders an intermediate three-dimensional r
epresentation of a scene in each denoising step. This enforces a strong inductiv
e structure within the diffusion process, providing a 3D consistent representati
on while only requiring 2D supervision. The resulting 3D representation can be r
endered from any view. We evaluate RenderDiffusion on FFHQ, AFHQ, ShapeNet and C
LEVR datasets, showing competitive performance for generation of 3D scenes and i
nference of 3D scenes from 2D images. Additionally, our diffusion-based approach
 allows us to use 2D inpainting to edit 3D scenes.
****
Adaptive Data-Free Quantization
Biao Qian, Yang Wang, Richang Hong, Meng Wang; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7960-7968
Data-free quantization (DFQ) recovers the performance of quantized network (Q) w
ithout the original data, but generates the fake sample via a generator (G) by l
earning from full-precision network (P), which, however, is totally independent
of Q, overlooking the adaptability of the knowledge from generated samples, i.e.
, informative or not to the learning process of Q, resulting into the overflow o
f generalization error. Building on this, several critical questions -- how to m
easure the sample adaptability to Q under varied bit-width scenarios? whether th
e largest adaptability is the best? how to generate the samples with adaptive ad
aptability to improve Q's generalization? To answer the above questions, in this
 paper, we propose an Adaptive Data-Free Quantization (AdaDFQ) method, which rev
isits DFQ from a zero-sum game perspective upon the sample adaptability between
two players -- a generator and a quantized network. Following this viewpoint, we
 further define the disagreement and agreement samples to form two boundaries, w
here the margin between two boundaries is optimized to adaptively regulate the a
daptability of generated samples to Q, so as to address the over-and-under fitti
ng issues. Our AdaDFQ reveals: 1) the largest adaptability is NOT the best for s
ample generation to benefit Q's generalization; 2) the knowledge of the generate
d sample should not be informative to Q only, but also related to the category a
nd distribution information of the training data for P. The theoretical and empi
rical analysis validate the advantages of AdaDFQ over the state-of-the-arts. Our
 code is available at https://github.com/hfutqian/AdaDFQ.
****
Neural Vector Fields: Implicit Representation by Explicit Learning
Xianghui Yang, Guosheng Lin, Zhenghao Chen, Luping Zhou; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 167
27-16738
Deep neural networks (DNNs) are widely applied for nowadays 3D surface reconstru
ction tasks and such methods can be further divided into two categories, which r
espectively warp templates explicitly by moving vertices or represent 3D surface
s implicitly as signed or unsigned distance functions. Taking advantage of both
advanced explicit learning process and powerful representation ability of implic
it functions, we propose a novel 3D representation method, Neural Vector Fields
(NVF). It not only adopts the explicit learning process to manipulate meshes dir

ectly, but also leverages the implicit representation of unsigned distance functions (UDFs) to break the barriers in resolution and topology. Specifically, our method first predicts the displacements from queries towards the surface and models the shapes as Vector Fields. Rather than relying on network differentiation to obtain direction fields as most existing UDF-based methods, the produced vector fields encode the distance and direction fields both and mitigate the ambiguity at "ridge" points, such that the calculation of direction fields is straightforward and differentiation-free. The differentiation-free characteristic enables us to further learn a shape codebook via Vector Quantization, which encodes the cross-object priors, accelerates the training procedure, and boosts model generalization on cross-category reconstruction. The extensive experiments on surface reconstruction benchmarks indicate that our method outperforms those state-of-the-art methods in different evaluation scenarios including watertight vs non-watertight shapes, category-specific vs category-agnostic reconstruction, category-unseen reconstruction, and cross-domain reconstruction. Our code is released at https://github.com/Wi-sc/NVF.
****
Latent-NeRF for Shape-Guided Generation of 3D Shapes and Textures
Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12663-12673
Text-guided image generation has progressed rapidly in recent years, inspiring major breakthroughs in text-guided shape generation. Recently, it has been shown that using score distillation, one can successfully text-guide a NeRF model to generate a 3D object. We adapt the score distillation to the publicly available, and computationally efficient, Latent Diffusion Models, which apply the entire diffusion process in a compact latent space of a pretrained autoencoder. As NeRFs operate in image space, a naive solution for guiding them with latent score distillation would require encoding to the latent space at each guidance step. Instead, we propose to bring the NeRF to the latent space, resulting in a Latent-NeRF. Analyzing our Latent-NeRF, we show that while Text-to-3D models can generate impressive results, they are inherently unconstrained and may lack the ability to guide or enforce a specific 3D structure. To assist and direct the 3D generation, we propose to guide our Latent-NeRF using a Sketch-Shape: an abstract geometry that defines the coarse structure of the desired object. Then, we present means to integrate such a constraint directly into a Latent-NeRF. This unique combination of text and shape guidance allows for increased control over the generation process. We also show that latent score distillation can be successfully applied directly on 3D meshes. This allows for generating high-quality textures on a given geometry. Our experiments validate the power of our different forms of guidance and the efficiency of using latent rendering.
****
Learning Generative Structure Prior for Blind Text Image Super-Resolution
Xiaoming Li, Wangmeng Zuo, Chen Change Loy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10103-10113
Blind text image super-resolution (SR) is challenging as one needs to cope with diverse font styles and unknown degradation. To address the problem, existing methods perform character recognition in parallel to regularize the SR task, either through a loss constraint or intermediate feature condition. Nonetheless, the high-level prior could still fail when encountering severe degradation. The problem is further compounded given characters of complex structures, e.g., Chinese characters that combine multiple pictographic or ideographic symbols into a single character. In this work, we present a novel prior that focuses more on the character structure. In particular, we learn to encapsulate rich and diverse structures in a StyleGAN and exploit such generative structure priors for restoration. To restrict the generative space of StyleGAN so that it obeys the structure of characters yet remains flexible in handling different font styles, we store the discrete features for each character in a codebook . The code subsequently drives the StyleGAN to generate high-resolution structural details to aid text SR. Compared to priors based on character recognition, the proposed structure prior

exerts stronger character-specific guidance to restore faithful and precise strokes of a designated character. Extensive experiments on synthetic and real datasets demonstrate the compelling performance of the proposed generative structure prior in facilitating robust text SR. Our code is available at https://github.com/csxmli2016/MARCONet.
****
Overcoming the Trade-Off Between Accuracy and Plausibility in 3D Hand Shape Reconstruction
Ziwei Yu, Chen Li, Linlin Yang, Xiaoxu Zheng, Michael Bi Mi, Gim Hee Lee, Angela Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 544-553
Direct mesh fitting for 3D hand shape reconstruction estimates highly accurate meshes. However, the resulting meshes are prone to artifacts and do not appear as plausible hand shapes. Conversely, parametric models like MANO ensure plausible hand shapes but are not as accurate as the non-parametric methods. In this work, we introduce a novel weakly-supervised hand shape estimation framework that integrates non-parametric mesh fitting with MANO models in an end-to-end fashion. Our joint model overcomes the tradeoff in accuracy and plausibility to yield well-aligned and high-quality 3D meshes, especially in challenging two-hand and hand-object interaction scenarios.
****
Open-Vocabulary Attribute Detection
María A. Bravo, Sudhanshu Mittal, Simon Ging, Thomas Brox; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7041-7050
Vision-language modeling has enabled open-vocabulary tasks where predictions can be queried using any text prompt in a zero-shot manner. Existing open-vocabulary tasks focus on object classes, whereas research on object attributes is limited due to the lack of a reliable attribute-focused evaluation benchmark. This paper introduces the Open-Vocabulary Attribute Detection (OVAD) task and the corresponding OVAD benchmark. The objective of the novel task and benchmark is to probe object-level attribute information learned by vision-language models. To this end, we created a clean and densely annotated test set covering 117 attribute classes on the 80 object classes of MS COCO. It includes positive and negative annotations, which enables open-vocabulary evaluation. Overall, the benchmark consists of 1.4 million annotations. For reference, we provide a first baseline method for open-vocabulary attribute detection. Moreover, we demonstrate the benchmark's value by studying the attribute detection performance of several foundation models.
****
PEFAT: Boosting Semi-Supervised Medical Image Classification via Pseudo-Loss Estimation and Feature Adversarial Training
Qingjie Zeng, Yutong Xie, Zilin Lu, Yong Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15671-15680
Pseudo-labeling approaches have been proven beneficial for semi-supervised learning (SSL) schemes in computer vision and medical imaging. Most works are dedicated to finding samples with high-confidence pseudo-labels from the perspective of model predicted probability. Whereas this way may lead to the inclusion of incorrectly pseudo-labeled data if the threshold is not carefully adjusted. In addition, low-confidence probability samples are frequently disregarded and not employed to their full potential. In this paper, we propose a novel Pseudo-loss Estimation and Feature Adversarial Training semi-supervised framework, termed as PEFAT, to boost the performance of multi-class and multi-label medical image classification from the point of loss distribution modeling and adversarial training. Specifically, we develop a trustworthy data selection scheme to split a high-quality pseudo-labeled set, inspired by the dividable pseudo-loss assumption that clean data tend to show lower loss while noise data is the opposite. Instead of directly discarding these samples with low-quality pseudo-labels, we present a novel regularization approach to learn discriminate information from them via injecting adversarial noises at the feature-level to smooth the decision boundary. Ex

perimental results on three medical and two natural image benchmarks validate that our PEFAT can achieve a promising performance and surpass other state-of-the-art methods. The code is available at https://github.com/maxwell0027/PEFAT.
****
TBP-Former: Learning Temporal Bird's-Eye-View Pyramid for Joint Perception and Prediction in Vision-Centric Autonomous Driving
Shaoheng Fang, Zi Wang, Yiqi Zhong, Junhao Ge, Siheng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1368-1378
Vision-centric joint perception and prediction (PnP) has become an emerging trend in autonomous driving research. It predicts the future states of the traffic participants in the surrounding environment from raw RGB images. However, it is still a critical challenge to synchronize features obtained at multiple camera views and timestamps due to inevitable geometric distortions and further exploit those spatial-temporal features. To address this issue, we propose a temporal bird's-eye-view pyramid transformer (TBP-Former) for vision-centric PnP, which includes two novel designs. First, a pose-synchronized BEV encoder is proposed to map raw image inputs with any camera pose at any time to a shared and synchronized BEV space for better spatial-temporal synchronization. Second, a spatial-temporal pyramid transformer is introduced to comprehensively extract multi-scale BEV features and predict future BEV states with the support of spatial priors. Extensive experiments on nuScenes dataset show that our proposed framework overall outperforms all state-of-the-art vision-based prediction methods.
****
Ground-Truth Free Meta-Learning for Deep Compressive Sampling
Xinran Qin, Yuhui Quan, Tongyao Pang, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9947-9956
Deep learning has become an important tool for reconstructing images in compressive sampling (CS). This paper proposes a ground-truth (GT) free meta-learning method for CS, which leverages both external and internal learning for unsupervised high-quality image reconstruction. The proposed method first trains a deep model via external meta-learning using only CS measurements, and then efficiently adapts the trained model to a test sample for further improvement by exploiting its internal characteristics. The meta-learning and model adaptation are built on an improved Stein's unbiased risk estimator (iSURE) that provides efficient computation and effective guidance for accurate prediction in the range space of the adjoint of the measurement matrix. To further improve the learning on the null space of the measurement matrix, a modified model-agnostic meta-learning scheme is proposed, along with a null-space-consistent loss and a bias-adaptive deep unrolling network to improve and accelerate model adaption in test time. Experimental results have demonstrated that the proposed GT-free method performs well, and can even compete with supervised learning-based methods.
****
SHS-Net: Learning Signed Hyper Surfaces for Oriented Normal Estimation of Point Clouds
Qing Li, Huifang Feng, Kanle Shi, Yue Gao, Yi Fang, Yu-Shen Liu, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13591-13600
We propose a novel method called SHS-Net for oriented normal estimation of point clouds by learning signed hyper surfaces, which can accurately predict normals with global consistent orientation from various point clouds. Almost all existing methods estimate oriented normals through a two-stage pipeline, i.e., unoriented normal estimation and normal orientation, and each step is implemented by a separate algorithm. However, previous methods are sensitive to parameter settings, resulting in poor results from point clouds with noise, density variations and complex geometries. In this work, we introduce signed hyper surfaces (SHS), which are parameterized by multi-layer perceptron (MLP) layers, to learn to estimate oriented normals from point clouds in an end-to-end manner. The signed hyper surfaces are implicitly learned in a high-dimensional feature space where the local and global information is aggregated. Specifically, we introduce a patch enco

ding module and a shape encoding module to encode a 3D point cloud into a local latent code and a global latent code, respectively. Then, an attention-weighted normal prediction module is proposed as a decoder, which takes the local and global latent codes as input to predict oriented normals. Experimental results show that our SHS-Net outperforms the state-of-the-art methods in both unoriented and oriented normal estimation on the widely used benchmarks. The code, data and pretrained models are available at https://github.com/LeoQLi/SHS-Net.
****
DistractFlow: Improving Optical Flow Estimation via Realistic Distractions and Pseudo-Labeling
Jisoo Jeong, Hong Cai, Risheek Garrepalli, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13691-13700
We propose a novel data augmentation approach, DistractFlow, for training optical flow estimation models by introducing realistic distractions to the input frames. Based on a mixing ratio, we combine one of the frames in the pair with a distractor image depicting a similar domain, which allows for inducing visual perturbations congruent with natural objects and scenes. We refer to such pairs as distracted pairs. Our intuition is that using semantically meaningful distractors enables the model to learn related variations and attain robustness against challenging deviations, compared to conventional augmentation schemes focusing only on low-level aspects and modifications. More specifically, in addition to the supervised loss computed between the estimated flow for the original pair and its ground-truth flow, we include a second supervised loss defined between the distracted pair's flow and the original pair's ground-truth flow, weighted with the same mixing ratio. Furthermore, when unlabeled data is available, we extend our augmentation approach to self-supervised settings through pseudo-labeling and cross-consistency regularization. Given an original pair and its distracted version, we enforce the estimated flow on the distracted pair to agree with the flow of the original pair. Our approach allows increasing the number of available training pairs significantly without requiring additional annotations. It is agnostic to the model architecture and can be applied to training any optical flow estimation models. Our extensive evaluations on multiple benchmarks, including Sintel, KITTI, and SlowFlow, show that DistractFlow improves existing models consistently, outperforming the latest state of the art.
****
Test of Time: Instilling Video-Language Models With a Sense of Time
Piyush Bagad, Makarand Tapaswi, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2503-2516
Modelling and understanding time remains a challenge in contemporary video understanding models. With language emerging as a key driver towards powerful generalization, it is imperative for foundational video-language models to have a sense of time. In this paper, we consider a specific aspect of temporal understanding: consistency of time order as elicited by before/after relations. We establish that seven existing video-language models struggle to understand even such simple temporal relations. We then question whether it is feasible to equip these foundational models with temporal awareness without re-training them from scratch. Towards this, we propose a temporal adaptation recipe on top of one such model, VideoCLIP, based on post-pretraining on a small amount of video-text data. We conduct a zero-shot evaluation of the adapted models on six datasets for three downstream tasks which require varying degrees of time awareness. We observe encouraging performance gains especially when the task needs higher time awareness. Our work serves as a first step towards probing and instilling a sense of time in existing video-language models without the need for data and compute-intense training from scratch.
****
Learning To Segment Every Referring Object Point by Point
Mengxue Qu, Yu Wu, Yunchao Wei, Wu Liu, Xiaodan Liang, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3021-3030

Referring Expression Segmentation (RES) can facilitate pixel-level semantic alig nment between vision and language. Most of the existing RES approaches require m assive pixel-level annotations, which are expensive and exhaustive. In this pape r, we propose a new partially supervised training paradigm for RES, i.e., traini ng using abundant referring bounding boxes and only a few (e.g., 1%) pixel-level referring masks. To maximize the transferability from the REC model, we constru ct our model based on the point-based sequence prediction model. We propose the co-content teacher-forcing to make the model explicitly associate the point coor dinates (scale values) with the referred spatial features, which alleviates the exposure bias caused by the limited segmentation masks. To make the most of refe rring bounding box annotations, we further propose the resampling pseudo points strategy to select more accurate pseudo-points as supervision. Extensive experim ents show that our model achieves 52.06% in terms of accuracy (versus 58.93% in fully supervised setting) on RefCOCO+@testA, when only using 1% of the mask anno tations.
****
Seeing With Sound: Long-range Acoustic Beamforming for Multimodal Scene Understa nding
Praneeth Chakravarthula, Jim Aldon D’Souza, Ethan Tseng, Joe Bartusek, Felix Hei de; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn ition (CVPR), 2023, pp. 982-991
Existing autonomous vehicles primarily use sensors that rely on electromagnetic waves which are undisturbed in good environmental conditions but can suffer in a dverse scenarios, such as low light or for objects with low reflectance. Moreove r, only objects in direct line-of-sight are typically detected by these existing methods. Acoustic pressure waves emanating from road users do not share these l imitations. However, such signals are typically ignored in automotive perception because they suffer from low spatial resolution and lack directional informatio n. In this work, we introduce long-range acoustic beamforming of pressure waves from noise directly produced by automotive vehicles in-the-wild as a complement ary sensing modality to traditional optical sensor approaches for detection of objects in dynamic traffic environments. To this end, we introduce the first mul timodal long-range acoustic beamforming dataset. We propose a neural aperture ex pansion method for beamforming and we validate its utility for multimodal automo tive object detection. We validate the benefit of adding sound detections to exi sting RGB cameras in challenging automotive scenarios, where camera-only approac hes fail or do not deliver the ultra-fast rates of pressure sensors.
****
OpenScene: 3D Scene Understanding With Open Vocabularies
Songyou Peng, Kyle Genova, Chiyu “Max” Jiang, Andrea Tagliasacchi, Marc Pollefey s, Thomas Funkhouser; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 815-824
Traditional 3D scene understanding approaches rely on labeled 3D datasets to tra in a model for a single task with supervision. We propose OpenScene, an alternat ive approach where a model predicts dense features for 3D scene points that are co-embedded with text and image pixels in CLIP feature space. This zero-shot app roach enables task-agnostic training and open-vocabulary queries. For example, t o perform SOTA zero-shot 3D semantic segmentation it first infers CLIP features for every 3D point and later classifies them based on similarities to embeddings of arbitrary class labels. More interestingly, it enables a suite of open-vocab ulary scene understanding applications that have never been done before. For exa mple, it allows a user to enter an arbitrary text query and then see a heat map indicating which parts of a scene match. Our approach is effective at identifyin g objects, materials, affordances, activities, and room types in complex 3D scen es, all using a single model trained without any labeled 3D data.
****
Movies2Scenes: Using Movie Metadata To Learn Scene Representation
Shixing Chen, Chun-Hao Liu, Xiang Hao, Xiaohan Nie, Maxim Arap, Raffay Hamid; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6535-6544

Understanding scenes in movies is crucial for a variety of applications such as video moderation, search, and recommendation. However, labeling individual scenes is a time-consuming process. In contrast, movie level metadata (e.g., genre, synopsis, etc.) regularly gets produced as part of the film production process, and is therefore significantly more commonly available. In this work, we propose a novel contrastive learning approach that uses movie metadata to learn a general-purpose scene representation. Specifically, we use movie metadata to define a measure of movie similarity, and use it during contrastive learning to limit our search for positive scene-pairs to only the movies that are considered similar to each other. Our learned scene representation consistently outperforms existing state-of-the-art methods on a diverse set of tasks evaluated using multiple benchmark datasets. Notably, our learned representation offers an average improvement of 7.9% on the seven classification tasks and 9.7% improvement on the two regression tasks in LVU dataset. Furthermore, using a newly collected movie dataset, we present comparative results of our scene representation on a set of video moderation tasks to demonstrate its generalizability on previously less explored tasks.
****
Think Twice Before Driving: Towards Scalable Decoders for End-to-End Autonomous Driving
Xiaosong Jia, Penghao Wu, Li Chen, Jiangwei Xie, Conghui He, Junchi Yan, Hongyang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21983-21994
End-to-end autonomous driving has made impressive progress in recent years. Existing methods usually adopt the decoupled encoder-decoder paradigm, where the encoder extracts hidden features from raw sensor data, and the decoder outputs the ego-vehicle's future trajectories or actions. Under such a paradigm, the encoder does not have access to the intended behavior of the ego agent, leaving the burden of finding out safety-critical regions from the massive receptive field and inferring about future situations to the decoder. Even worse, the decoder is usually composed of several simple multi-layer perceptrons (MLP) or GRUs while the encoder is delicately designed (e.g., a combination of heavy ResNets or Transformer). Such an imbalanced resource-task division hampers the learning process. In this work, we aim to alleviate the aforementioned problem by two principles: (1) fully utilizing the capacity of the encoder; (2) increasing the capacity of the decoder. Concretely, we first predict a coarse-grained future position and action based on the encoder features. Then, conditioned on the position and action, the future scene is imagined to check the ramification if we drive accordingly. We also retrieve the encoder features around the predicted coordinate to obtain fine-grained information about the safety-critical region. Finally, based on the predicted future and the retrieved salient feature, we refine the coarse-grained position and action by predicting its offset from ground-truth. The above refinement module could be stacked in a cascaded fashion, which extends the capacity of the decoder with spatial-temporal prior knowledge about the conditioned future. We conduct experiments on the CARLA simulator and achieve state-of-the-art performance in closed-loop benchmarks. Extensive ablation studies demonstrate the effectiveness of each proposed module. Code and models are available at https://github.com/opendrivelab/ThinkTwice.
****
DSVT: Dynamic Sparse Voxel Transformer With Rotated Sets
Haiyang Wang, Chen Shi, Shaoshuai Shi, Meng Lei, Sen Wang, Di He, Bernt Schiele, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13520-13529
Designing an efficient yet deployment-friendly 3D backbone to handle sparse point clouds is a fundamental problem in 3D perception. Compared with the customized sparse convolution, the attention mechanism in Transformers is more appropriate for flexibly modeling long-range relationships and is easier to be deployed in real-world applications. However, due to the sparse characteristics of point clouds, it is non-trivial to apply a standard transformer on sparse points. In this paper, we present Dynamic Sparse Voxel Transformer (DSVT), a single-stride wind

ow-based voxel Transformer backbone for outdoor 3D perception. In order to effic iently process sparse points in parallel, we propose Dynamic Sparse Window Atten tion, which partitions a series of local regions in each window according to its sparsity and then computes the features of all regions in a fully parallel mann er. To allow the cross-set connection, we design a rotated set partitioning stra tegy that alternates between two partitioning configurations in consecutive self -attention layers. To support effective downsampling and better encode geometric information, we also propose an attention-style 3D pooling module on sparse poi nts, which is powerful and deployment-friendly without utilizing any customized CUDA operations. Our model achieves state-of-the-art performance with a broad ra nge of 3D perception tasks. More importantly, DSVT can be easily deployed by Ten sorRT with real-time inference speed (27Hz). Code will be available at https://g ithub.com/Haiyang-W/DSVT.
****
Joint Token Pruning and Squeezing Towards More Aggressive Compression of Vision Transformers
Siyuan Wei, Tianzhu Ye, Shen Zhang, Yao Tang, Jiajun Liang; Proceedings of the I EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2092-2101
Although vision transformers (ViTs) have shown promising results in various comp uter vision tasks recently, their high computational cost limits their practical applications. Previous approaches that prune redundant tokens have demonstrated a good trade-off between performance and computation costs. Nevertheless, error s caused by pruning strategies can lead to significant information loss. Our qua ntitative experiments reveal that the impact of pruned tokens on performance sho uld be noticeable. To address this issue, we propose a novel joint Token Pruning & Squeezing module (TPS) for compressing vision transformers with higher effici ency. Firstly, TPS adopts pruning to get the reserved and pruned subsets. Second ly, TPS squeezes the information of pruned tokens into partial reserved tokens v ia the unidirectional nearest-neighbor matching and similarity-oriented fusing s teps. Compared to state-of-the-art methods, our approach outperforms them under all token pruning intensities. Especially while shrinking DeiT-tiny&small comput ational budgets to 35%, it improves the accuracy by 1%-6% compared with baseline s on ImageNet classification. The proposed method can accelerate the throughput of DeiT-small beyond DeiT-tiny, while its accuracy surpasses DeiT-tiny by 4.78%. Experiments on various transformers demonstrate the effectiveness of our method , while analysis experiments prove our higher robustness to the errors of the to ken pruning policy. Code is available at https://github.com/megvii-research/TPS-CVPR2023.
****
Enhancing the Self-Universality for Transferable Targeted Attacks
Zhipeng Wei, Jingjing Chen, Zuxuan Wu, Yu-Gang Jiang; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12281-12290
In this paper, we propose a novel transfer-based targeted attack method that opt imizes the adversarial perturbations without any extra training efforts for auxi liary networks on training data. Our new attack method is proposed based on the observation that highly universal adversarial perturbations tend to be more tran sferable for targeted attacks. Therefore, we propose to make the perturbation to be agnostic to different local regions within one image, which we called as sel f-universality. Instead of optimizing the perturbations on different images, opt imizing on different regions to achieve self-universality can get rid of using e xtra data. Specifically, we introduce a feature similarity loss that encourages the learned perturbations to be universal by maximizing the feature similarity b etween adversarial perturbed global images and randomly cropped local regions. W ith the feature similarity loss, our method makes the features from adversarial perturbations to be more dominant than that of benign images, hence improving ta rgeted transferability. We name the proposed attack method as Self-Universality (SU) attack. Extensive experiments demonstrate that SU can achieve high success rates for transfer-based targeted attacks. On ImageNet-compatible dataset, SU yi

elds an improvement of 12% compared with existing state-of-the-art methods. Code is available at https://github.com/zhipeng-wei/Self-Universality.
****
Disentangling Orthogonal Planes for Indoor Panoramic Room Layout Estimation With Cross-Scale Distortion Awareness
Zhijie Shen, Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Shuai Zheng, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17337-17345
Based on the Manhattan World assumption, most existing indoor layout estimation schemes focus on recovering layouts from vertically compressed 1D sequences. However, the compression procedure confuses the semantics of different planes, yielding inferior performance with ambiguous interpretability. To address this issue, we propose to disentangle this 1D representation by pre-segmenting orthogonal (vertical and horizontal) planes from a complex scene, explicitly capturing the geometric cues for indoor layout estimation. Considering the symmetry between the floor boundary and ceiling boundary, we also design a soft-flipping fusion strategy to assist the pre-segmentation. Besides, we present a feature assembling mechanism to effectively integrate shallow and deep features with distortion distribution awareness. To compensate for the potential errors in pre-segmentation, we further leverage triple attention to reconstruct the disentangled sequences for better performance. Experiments on four popular benchmarks demonstrate our superiority over existing SoTA solutions, especially on the 3DIoU metric. The code is available at https://github.com/zhijieshen-bjtu/DOPNet.
****
EditableNeRF: Editing Topologically Varying Neural Radiance Fields by Key Points
Chengwei Zheng, Wenbin Lin, Feng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8317-8327
Neural radiance fields (NeRF) achieve highly photo-realistic novel-view synthesis, but it's a challenging problem to edit the scenes modeled by NeRF-based methods, especially for dynamic scenes. We propose editable neural radiance fields that enable end-users to easily edit dynamic scenes and even support topological changes. Input with an image sequence from a single camera, our network is trained fully automatically and models topologically varying dynamics using our picked-out surface key points. Then end-users can edit the scene by easily dragging the key points to desired new positions. To achieve this, we propose a scene analysis method to detect and initialize key points by considering the dynamics in the scene, and a weighted key points strategy to model topologically varying dynamics by joint key points and weights optimization. Our method supports intuitive multi-dimensional (up to 3D) editing and can generate novel scenes that are unseen in the input sequence. Experiments demonstrate that our method achieves high-quality editing on various dynamic scenes and outperforms the state-of-the-art. Our code and captured data are available at https://chengwei-zheng.github.io/EditableNeRF/.
****
Neural Map Prior for Autonomous Driving
Xuan Xiong, Yicheng Liu, Tianyuan Yuan, Yue Wang, Yilun Wang, Hang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17535-17544
High-definition (HD) semantic maps are a crucial component for autonomous driving on urban streets. Traditional offline HD maps are created through labor-intensive manual annotation processes, which are costly and do not accommodate timely updates. Recently, researchers have proposed to infer local maps based on online sensor observations. However, the range of online map inference is constrained by sensor perception range and is easily affected by occlusions. In this work, we propose Neural Map Prior (NMP), a neural representation of global maps that enables automatic global map updates and enhances local map inference performance. To incorporate the strong map prior into local map inference, we leverage cross-attention to dynamically capture the correlations between current features and prior features. For updating the global neural map prior, we use a learning-based fusion module to guide the network in fusing features from previous traversals

. This design allows the network to capture a global neural map prior while making sequential online map predictions. Experimental results on the nuScenes dataset demonstrate that our framework is compatible with most map segmentation/detection methods, improving map prediction performance in challenging weather conditions and over an extended horizon. To the best of our knowledge, this represents the first learning-based system for constructing a global map prior.
****
Solving Oscillation Problem in Post-Training Quantization Through a Theoretical Perspective
Yuexiao Ma, Huixia Li, Xiawu Zheng, Xuefeng Xiao, Rui Wang, Shilei Wen, Xin Pan, Fei Chao, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7950-7959
Post-training quantization (PTQ) is widely regarded as one of the most efficient compression methods practically, benefitting from its data privacy and low computation costs. We argue that an overlooked problem of oscillation is in the PTQ methods. In this paper, we take the initiative to explore and present a theoretical proof to explain why such a problem is essential in PTQ. And then, we try to solve this problem by introducing a principled and generalized framework theoretically. In particular, we first formulate the oscillation in PTQ and prove the problem is caused by the difference in module capacity. To this end, we define the module capacity (ModCap) under data-dependent and data-free scenarios, where the differentials between adjacent modules are used to measure the degree of oscillation. The problem is then solved by selecting top-k differentials, in which the corresponding modules are jointly optimized and quantized. Extensive experiments demonstrate that our method successfully reduces the performance drop and is generalized to different neural networks and PTQ methods. For example, with 2/4 bit ResNet-50 quantization, our method surpasses the previous state-of-the-art method by 1.9%. It becomes more significant on small model quantization, e.g. surpasses BRECQ method by 6.61% on MobileNetV2*0.5.
****
PEAL: Prior-Embedded Explicit Attention Learning for Low-Overlap Point Cloud Registration
Junle Yu, Luwei Ren, Yu Zhang, Wenhui Zhou, Lili Lin, Guojun Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17702-17711
Learning distinctive point-wise features is critical for low-overlap point cloud registration. Recently, it has achieved huge success in incorporating Transformer into point cloud feature representation, which usually adopts a self-attention module to learn intra-point-cloud features first, then utilizes a cross-attention module to perform feature exchange between input point clouds. Self-attention is computed by capturing the global dependency in geometric space. However, this global dependency can be ambiguous and lacks distinctiveness, especially in indoor low-overlap scenarios, as which the dependence with an extensive range of non-overlapping points introduces ambiguity. To address this issue, we present PEAL, a Prior-embedded Explicit Attention Learning model. By incorporating prior knowledge into the learning process, the points are divided into two parts. One includes points lying in the putative overlapping region and the other includes points lying in the putative non-overlapping region. Then PEAL explicitly learns one-way attention with the putative overlapping points. This simplistic design attains surprising performance, significantly relieving the aforementioned feature ambiguity. Our method improves the Registration Recall by 6+% on the challenging 3DLoMatch benchmark and achieves state-of-the-art performance on Feature Matching Recall, Inlier Ratio, and Registration Recall on both 3DMatch and 3DLoMatch. Code will be made publicly available.
****
NeuralEditor: Editing Neural Radiance Fields via Manipulating Point Clouds
Jun-Kun Chen, Jipeng Lyu, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12439-12448
This paper proposes NeuralEditor that enables neural radiance fields (NeRFs) natively editable for general shape editing tasks. Despite their impressive results

 on novel-view synthesis, it remains a fundamental challenge for NeRFs to edit t
he shape of the scene. Our key insight is to exploit the explicit point cloud re
presentation as the underlying structure to construct NeRFs, inspired by the int
uitive interpretation of NeRF rendering as a process that projects or "plots" th
e associated 3D point cloud to a 2D image plane. To this end, NeuralEditor intro
duces a novel rendering scheme based on deterministic integration within K-D tre
e-guided density-adaptive voxels, which produces both high-quality rendering res
ults and precise point clouds through optimization. NeuralEditor then performs s
hape editing via mapping associated points between point clouds. Extensive evalu
ation shows that NeuralEditor achieves state-of-the-art performance in both shap
e deformation and scene morphing tasks. Notably, NeuralEditor supports both zero
-shot inference and further fine-tuning over the edited scene. Our code, benchma
rk, and demo video are available at https://immortalco.github.io/NeuralEditor.
****
NIKI: Neural Inverse Kinematics With Invertible Neural Networks for 3D Human Pos
e and Shape Estimation
Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, Cewu Lu; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
3, pp. 12933-12942
With the progress of 3D human pose and shape estimation, state-of-the-art method
s can either be robust to occlusions or obtain pixel-aligned accuracy in non-occ
lusion cases. However, they cannot obtain robustness and mesh-image alignment at
 the same time. In this work, we present NIKI (Neural Inverse Kinematics with In
vertible Neural Network), which models bi-directional errors to improve the robu
stness to occlusions and obtain pixel-aligned accuracy. NIKI can learn from both
 the forward and inverse processes with invertible networks. In the inverse proc
ess, the model separates the error from the plausible 3D pose manifold for a rob
ust 3D human pose estimation. In the forward process, we enforce the zero-error
boundary conditions to improve the sensitivity to reliable joint positions for b
etter mesh-image alignment. Furthermore, NIKI emulates the analytical inverse ki
nematics algorithms with the twist-and-swing decomposition for better interpreta
bility. Experiments on standard and occlusion-specific benchmarks demonstrate th
e effectiveness of NIKI, where we exhibit robust and well-aligned results simult
aneously. Code is available at https://github.com/Jeff-sjtu/NIKI
****
Masked Image Modeling With Local Multi-Scale Reconstruction
Haoqing Wang, Yehui Tang, Yunhe Wang, Jianyuan Guo, Zhi-Hong Deng, Kai Han; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2023, pp. 2122-2131
Masked Image Modeling (MIM) achieves outstanding success in self-supervised repr
esentation learning. Unfortunately, MIM models typically have huge computational
 burden and slow learning process, which is an inevitable obstacle for their ind
ustrial applications. Although the lower layers play the key role in MIM, existi
ng MIM models conduct reconstruction task only at the top layer of encoder. The
lower layers are not explicitly guided and the interaction among their patches i
s only used for calculating new activations. Considering the reconstruction task
 requires non-trivial inter-patch interactions to reason target signals, we appl
y it to multiple local layers including lower and upper layers. Further, since t
he multiple layers expect to learn the information of different scales, we desig
n local multi-scale reconstruction, where the lower and upper layers reconstruct
 fine-scale and coarse-scale supervision signals respectively. This design not o
nly accelerates the representation learning process by explicitly guiding multip
le layers, but also facilitates multi-scale semantical understanding to the inpu
t. Extensive experiments show that with significantly less pre-training burden,
our model achieves comparable or better performance on classification, detection
 and segmentation tasks than existing MIM models.
****
Transfer4D: A Framework for Frugal Motion Capture and Deformation Transfer
Shubh Maheshwari, Rahul Narain, Ramya Hebbalaguppe; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12836-12

Animating a virtual character based on a real performance of an actor is a chall
enging task that currently requires expensive motion capture setups and addition
al effort by expert animators, rendering it accessible only to large production
houses. The goal of our work is to democratize this task by developing a frugal
alternative termed "Transfer4D" that uses only commodity depth sensors and furth
er reduces animators' effort by automating the rigging and animation transfer pr
ocess. To handle sparse, incomplete videos from depth video inputs and large var
iations between source and target objects, we propose to use skeletons as an int
ermediary representation between motion capture and transfer. We propose a novel
skeleton extraction pipeline from single-view depth sequence that incorporates
additional geometric information, resulting in superior performance in motion re
construction and transfer in comparison to the contemporary methods. We use non-
rigid reconstruction to track motion from the depth sequence, and then we rig th
e source object using skinning decomposition. Finally, the rig is embedded into
the target object for motion retargeting.
****
GeoVLN: Learning Geometry-Enhanced Visual Representation With Slot Attention for
Vision-and-Language Navigation
Jingyang Huo, Qiang Sun, Boyan Jiang, Haitao Lin, Yanwei Fu; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
23212-23221
Most existing works solving Room-to-Room VLN problem only utilize RGB images and
do not consider local context around candidate views, which lack sufficient vis
ual cues about surrounding environment. Moreover, natural language contains comp
lex semantic information thus its correlations with visual inputs are hard to mo
del merely with cross attention. In this paper, we propose GeoVLN, which learns
Geometry-enhanced visual representation based on slot attention for robust Visua
l-and-Language Navigation. The RGB images are compensated with the corresponding
depth maps and normal maps predicted by Omnidata as visual inputs. Technically,
we introduce a two-stage module that combine local slot attention and CLIP mode
l to produce geometry-enhanced representation from such input. We employ V&L BER
T to learn a cross-modal representation that incorporate both language and visio
n informations. Additionally, a novel multiway attention module is designed, enc
ouraging different phrases of input instruction to exploit the most related feat
ures from visual input. Extensive experiments demonstrate the effectiveness of o
ur newly designed modules and show the compelling performance of the proposed me
thod.
****
KiUT: Knowledge-Injected U-Transformer for Radiology Report Generation
Zhongzhen Huang, Xiaofan Zhang, Shaoting Zhang; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19809-19818
Radiology report generation aims to automatically generate a clinically accurate
and coherent paragraph from the X-ray image, which could relieve radiologists f
rom the heavy burden of report writing. Although various image caption methods h
ave shown remarkable performance in the natural image field, generating accurate
reports for medical images requires knowledge of multiple modalities, including
vision, language, and medical terminology. We propose a Knowledge-injected U-Tr
ansformer (KiUT) to learn multi-level visual representation and adaptively disti
ll the information with contextual and clinical knowledge for word prediction. I
n detail, a U-connection schema between the encoder and decoder is designed to m
odel interactions between different modalities. And a symptom graph and an injec
ted knowledge distiller are developed to assist the report generation. Experimen
tally, we outperform state-of-the-art methods on two widely used benchmark datas
ets: IU-Xray and MIMIC-CXR. Further experimental results prove the advantages of
our architecture and the complementary benefits of the injected knowledge.
****
Flexible-Cm GAN: Towards Precise 3D Dose Prediction in Radiotherapy
Riqiang Gao, Bin Lou, Zhoubing Xu, Dorin Comaniciu, Ali Kamen; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p

p. 715-725
Deep learning has been utilized in knowledge-based radiotherapy planning in whic
h a system trained with a set of clinically approved plans is employed to infer
a three-dimensional dose map for a given new patient. However, previous deep met
hods are primarily limited to simple scenarios, e.g., a fixed planning type or a
 consistent beam angle configuration. This in fact limits the usability of such
approaches and makes them not generalizable over a larger set of clinical scenar
ios. Herein, we propose a novel conditional generative model, Flexible-C^m GAN,
utilizing additional information regarding planning types and various beam geome
tries. A miss-consistency loss is proposed to deal with the challenge of having
a limited set of conditions on the input data, e.g., incomplete training samples
. To address the challenges of including clinical preferences, we derive a diffe
rentiable shift-dose-volume loss to incorporate the well-known dose-volume histo
gram constraints. During inference, users can flexibly choose a specific plannin
g type and a set of beam angles to meet the clinical requirements. We conduct ex
periments on an illustrative face dataset to show the motivation of Flexible-C^m
 GAN and further validate our model's potential clinical values with two radioth
erapy datasets. The results demonstrate the superior performance of the proposed
 method in a practical heterogeneous radiotherapy planning application compared
to existing deep learning-based approaches.
****
Randomized Adversarial Training via Taylor Expansion
Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, Xiaowei Huang; Proceedings of the
 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp
. 16447-16457
In recent years, there has been an explosion of research into developing more ro
bust deep neural networks against adversarial examples. Adversarial training app
ears as one of the most successful methods. To deal with both the robustness aga
inst adversarial examples and the accuracy over clean examples, many works devel
op enhanced adversarial training methods to achieve various trade-offs between t
hem. Leveraging over the studies that smoothed update on weights during training
 may help find flat minima and improve generalization, we suggest reconciling th
e robustness-accuracy trade-off from another perspective, i.e., by adding random
 noise into deterministic weights. The randomized weights enable our design of a
 novel adversarial training method via Taylor expansion of a small Gaussian nois
e, and we show that the new adversarial training method can flatten loss landsca
pe and find flat minima. With PGD, CW, and Auto Attacks, an extensive set of exp
eriments demonstrate that our method enhances the state-of-the-art adversarial t
raining methods, boosting both robustness and clean accuracy. The code is availa
ble at https://github.com/Alexkael/Randomized-Adversarial-Training.
****
Handy: Towards a High Fidelity 3D Hand Shape and Appearance Model
Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileio
s Triantafyllou, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2023, pp. 4670-4680
Over the last few years, with the advent of virtual and augmented reality, an en
ormous amount of research has been focused on modeling, tracking and reconstruct
ing human hands. Given their power to express human behavior, hands have been a
very important, but challenging component of the human body. Currently, most of
the state-of-the-art reconstruction and pose estimation methods rely on the low
polygon MANO model. Apart from its low polygon count, MANO model was trained wit
h only 31 adult subjects, which not only limits its expressive power but also im
poses unnecessary shape reconstruction constraints on pose estimation methods. M
oreover, hand appearance remains almost unexplored and neglected from the majori
ty of hand reconstruction methods. In this work, we propose "Handy", a large-sca
le model of the human hand, modeling both shape and appearance composed of over
1200 subjects which we make publicly available for the benefit of the research c
ommunity. In contrast to current models, our proposed hand model was trained on
a dataset with large diversity in age, gender, and ethnicity, which tackles the
limitations of MANO and accurately reconstructs out-of-distribution samples. In

order to create a high quality texture model, we trained a powerful GAN, which preserves high frequency details and is able to generate high resolution hand textures. To showcase the capabilities of the proposed model, we built a synthetic dataset of textured hands and trained a hand pose estimation network to reconstruct both the shape and appearance from single images. As it is demonstrated in an extensive series of quantitative as well as qualitative experiments, our model proves to be robust against the state-of-the-art and realistically captures the 3D hand shape and pose along with a high frequency detailed texture even in adverse "in-the-wild" conditions.
****
Learning To Measure the Point Cloud Reconstruction Loss in a Representation Space
Tianxin Huang, Zhonggan Ding, Jiangning Zhang, Ying Tai, Zhenyu Zhang, Mingang Chen, Chengjie Wang, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12208-12217
For point cloud reconstruction-related tasks, the reconstruction losses to evaluate the shape differences between reconstructed results and the ground truths are typically used to train the task networks. Most existing works measure the training loss with point-to-point distance, which may introduce extra defects as predefined matching rules may deviate from the real shape differences. Although some learning-based works have been proposed to overcome the weaknesses of manually-defined rules, they still measure the shape differences in 3D Euclidean space, which may limit their ability to capture defects in reconstructed shapes. In this work, we propose a learning-based Contrastive Adversarial Loss (CALoss) to measure the point cloud reconstruction loss dynamically in a non-linear representation space by combining the contrastive constraint with the adversarial strategy. Specifically, we use the contrastive constraint to help CALoss learn a representation space with shape similarity, while we introduce the adversarial strategy to help CALoss mine differences between reconstructed results and ground truths. According to experiments on reconstruction-related tasks, CALoss can help task networks improve reconstruction performances and learn more representative representations.
****
Progressive Neighbor Consistency Mining for Correspondence Pruning
Xin Liu, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9527-9537
The goal of correspondence pruning is to recognize correct correspondences (inliers) from initial ones, with applications to various feature matching based tasks. Seeking neighbors in the coordinate and feature spaces is a common strategy in many previous methods. However, it is difficult to ensure that these neighbors are always consistent, since the distribution of false correspondences is extremely irregular. For addressing this problem, we propose a novel global-graph space to search for consistent neighbors based on a weighted global graph that can explicitly explore long-range dependencies among correspondences. On top of that, we progressively construct three neighbor embeddings according to different neighbor search spaces, and design a Neighbor Consistency block to extract neighbor context and explore their interactions sequentially. In the end, we develop a Neighbor Consistency Mining Network (NCMNet) for accurately recovering camera poses and identifying inliers. Experimental results indicate that our NCMNet achieves a significant performance advantage over state-of-the-art competitors on challenging outdoor and indoor matching scenes. The source code can be found at https://github.com/xinliu29/NCMNet.
****
Learning To Zoom and Unzoom
Chittesh Thavamani, Mengtian Li, Francesco Ferroni, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5086-5095
Many perception systems in mobile computing, autonomous navigation, and AR/VR face strict compute constraints that are particularly challenging for high-resolution input images. Previous works propose nonuniform downsamplers that "learn to

zoom" on salient image regions, reducing compute while retaining task-relevant i
mage information. However, for tasks with spatial labels (such as 2D/3D object d
etection and semantic segmentation), such distortions may harm performance. In t
his work (LZU), we "learn to zoom" in on the input image, compute spatial featur
es, and then "unzoom" to revert any deformations. To enable efficient and differ
entiable unzooming, we approximate the zooming warp with a piecewise bilinear ma
pping that is invertible. LZU can be applied to any task with 2D spatial input a
nd any model with 2D spatial features, and we demonstrate this versatility by ev
aluating on a variety of tasks and datasets: object detection on Argoverse-HD, s
emantic segmentation on Cityscapes, and monocular 3D object detection on nuScene
s. Interestingly, we observe boosts in performance even when high-resolution sen
sor data is unavailable, implying that LZU can be used to "learn to upsample" as
 well. Code and additional visuals are available at https://tchittesh.github.io/
lzu/.
****
Task Difficulty Aware Parameter Allocation & Regularization for Lifelong Learnin
g
Wenjin Wang, Yunqing Hu, Qianglong Chen, Yin Zhang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7776-778
5
Parameter regularization or allocation methods are effective in overcoming catas
trophic forgetting in lifelong learning. However, they solve all tasks in a sequ
ence uniformly and ignore the differences in the learning difficulty of differen
t tasks. So parameter regularization methods face significant forgetting when le
arning a new task very different from learned tasks, and parameter allocation me
thods face unnecessary parameter overhead when learning simple tasks. In this pa
per, we propose the Parameter Allocation & Regularization (PAR), which adaptivel
y select an appropriate strategy for each task from parameter allocation and reg
ularization based on its learning difficulty. A task is easy for a model that ha
s learned tasks related to it and vice versa. We propose a divergence estimation
 method based on the Nearest-Prototype distance to measure the task relatedness
using only features of the new task. Moreover, we propose a time-efficient relat
edness-aware sampling-based architecture search strategy to reduce the parameter
 overhead for allocation. Experimental results on multiple benchmarks demonstrat
e that, compared with SOTAs, our method is scalable and significantly reduces th
e model's redundancy while improving the model's performance. Further qualitativ
e analysis indicates that PAR obtains reasonable task-relatedness.
****
Bootstrapping Objectness From Videos by Relaxed Common Fate and Visual Grouping
Long Lian, Zhirong Wu, Stella X. Yu; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2023, pp. 14582-14591
We study learning object segmentation from unlabeled videos. Humans can easily s
egment moving objects without knowing what they are. The Gestalt law of common f
ate, i.e., what move at the same speed belong together, has inspired unsupervise
d object discovery based on motion segmentation. However, common fate is not a r
eliable indicator of objectness: Parts of an articulated / deformable object may
 not move at the same speed, whereas shadows / reflections of an object always m
ove with it but are not part of it. Our insight is to bootstrap objectness by fi
rst learning image features from relaxed common fate and then refining them base
d on visual appearance grouping within the image itself and across images statis
tically. Specifically, we learn an image segmenter first in the loop of approxim
ating optical flow with constant segment flow plus small within-segment residual
 flow, and then by refining it for more coherent appearance and statistical figu
re-ground relevance. On unsupervised video object segmentation, using only ResNe
t and convolutional heads, our model surpasses the state-of-the-art by absolute
gains of 7/9/5% on DAVIS16 / STv2 / FBMS59 respectively, demonstrating the effec
tiveness of our ideas. Our code is publicly available.
****
From Node Interaction To Hop Interaction: New Effective and Scalable Graph Learn
ing Paradigm

Jie Chen, Zilong Li, Yin Zhu, Junping Zhang, Jian Pu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7876-7885
Existing Graph Neural Networks (GNNs) follow the message-passing mechanism that conducts information interaction among nodes iteratively. While considerable progress has been made, such node interaction paradigms still have the following limitation. First, the scalability limitation precludes the broad application of GNNs in large-scale industrial settings since the node interaction among rapidly expanding neighbors incurs high computation and memory costs. Second, the over-smoothing problem restricts the discrimination ability of nodes, i.e., node representations of different classes will converge to indistinguishable after repeated node interactions. In this work, we propose a novel hop interaction paradigm to address these limitations simultaneously. The core idea is to convert the interaction target among nodes to pre-processed multi-hop features inside each node. We design a simple yet effective HopGNN framework that can easily utilize existing GNNs to achieve hop interaction. Furthermore, we propose a multi-task learning strategy with a self-supervised learning objective to enhance HopGNN. We conduct extensive experiments on 12 benchmark datasets in a wide range of domains, scales, and smoothness of graphs. Experimental results show that our methods achieve superior performance while maintaining high scalability and efficiency. The code is at https://github.com/JC-202/HopGNN.
****
Semi-Supervised Hand Appearance Recovery via Structure Disentanglement and Dual Adversarial Discrimination
Zimeng Zhao, Binghui Zuo, Zhiyu Long, Yangang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12125-12136
Enormous hand images with reliable annotations are collected through marker-based MoCap. Unfortunately, degradations caused by markers limit their application in hand appearance reconstruction. A clear appearance recovery insight is an image-to-image translation trained with unpaired data. However, most frameworks fail because there exists structure inconsistency from a degraded hand to a bare one. The core of our approach is to first disentangle the bare hand structure from those degraded images and then wrap the appearance to this structure with a dual adversarial discrimination (DAD) scheme. Both modules take full advantage of the semi-supervised learning paradigm: The structure disentanglement benefits from the modeling ability of ViT, and the translator is enhanced by the dual discrimination on both translation processes and translation results. Comprehensive evaluations have been conducted to prove that our framework can robustly recover photo-realistic hand appearance from diverse marker-contained and even object-occluded datasets. It provides a novel avenue to acquire bare hand appearance data for other downstream learning problems.
****
Understanding and Improving Features Learned in Deep Functional Maps
Souhaib Attaiki, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1316-1326
Deep functional maps have recently emerged as a successful paradigm for non-rigid 3D shape correspondence tasks. An essential step in this pipeline consists in learning feature functions that are used as constraints to solve for a functional map inside the network. However, the precise nature of the information learned and stored in these functions is not yet well understood. Specifically, a major question is whether these features can be used for any other objective, apart from their purely algebraic role, in solving for functional map matrices. In this paper, we show that under some mild conditions, the features learned within deep functional map approaches can be used as point-wise descriptors and thus are directly comparable across different shapes, even without the necessity of solving for a functional map at test time. Furthermore, informed by our analysis, we propose effective modifications to the standard deep functional map pipeline, which promotes structural properties of learned features, significantly improving the matching results. Finally, we demonstrate that previously unsuccessful attemp

ts at using extrinsic architectures for deep functional map feature extraction can be remedied via simple architectural changes, which promote the theoretical properties suggested by our analysis. We thus bridge the gap between intrinsic and extrinsic surface-based learning, suggesting the necessary and sufficient conditions for successful shape matching. Our code is available at https://github.com/pvnieo/clover.
****
Back to the Source: Diffusion-Driven Adaptation To Test-Time Corruption
Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, Dequan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11786-11796
Test-time adaptation harnesses test inputs to improve the accuracy of a model trained on source data when tested on shifted target data. Most methods update the source model by (re-)training on each target domain. While re-training can help, it is sensitive to the amount and order of the data and the hyperparameters for optimization. We update the target data instead, and project all test inputs toward the source domain with a generative diffusion model. Our diffusion-driven adaptation (DDA) method shares its models for classification and generation across all domains, training both on source then freezing them for all targets, to avoid expensive domain-wise re-training. We augment diffusion with image guidance and classifier self-ensembling to automatically decide how much to adapt. Input adaptation by DDA is more robust than model adaptation across a variety of corruptions, models, and data regimes on the ImageNet-C benchmark. With its input-wise updates, DDA succeeds where model adaptation degrades on too little data (small batches), on dependent data (correlated orders), or on mixed data (multiple corruptions).
****
PartManip: Learning Cross-Category Generalizable Part Manipulation Policy From Point Cloud Observations
Haoran Geng, Ziming Li, Yiran Geng, Jiayi Chen, Hao Dong, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2978-2988
Learning a generalizable object manipulation policy is vital for an embodied agent to work in complex real-world scenes. Parts, as the shared components in different object categories, have the potential to increase the generalization ability of the manipulation policy and achieve cross-category object manipulation. In this work, we build the first large-scale, part-based cross-category object manipulation benchmark, PartManip, which is composed of 11 object categories, 494 objects, and 1432 tasks in 6 task classes. Compared to previous work, our benchmark is also more diverse and realistic, i.e., having more objects and using sparse-view point cloud as input without oracle information like part segmentation. To tackle the difficulties of vision-based policy learning, we first train a state-based expert with our proposed part-based canonicalization and part-aware rewards, and then distill the knowledge to a vision-based student. We also find an expressive backbone is essential to overcome the large diversity of different objects. For cross-category generalization, we introduce domain adversarial learning for domain-invariant feature extraction. Extensive experiments in simulation show that our learned policy can outperform other methods by a large margin, especially on unseen object categories. We also demonstrate our method can successfully manipulate novel objects in the real world.
****
Polynomial Implicit Neural Representations for Large Diverse Datasets
Rajhans Singh, Ankita Shukla, Pavan Turaga; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2041-2051
Implicit neural representations (INR) have gained significant popularity for signal and image representation for many end-tasks, such as superresolution, 3D modeling, and more. Most INR architectures rely on sinusoidal positional encoding, which accounts for high-frequency information in data. However, the finite encoding size restricts the model's representational power. Higher representational power is needed to go from representing a single given image to representing larg

e and diverse datasets. Our approach addresses this gap by representing an image with a polynomial function and eliminates the need for positional encodings. Therefore, to achieve a progressively higher degree of polynomial representation, we use element-wise multiplications between features and affine-transformed coordinate locations after every ReLU layer. The proposed method is evaluated qualitatively and quantitatively on large datasets like ImageNet. The proposed Poly-INR model performs comparably to state-of-the-art generative models without any convolution, normalization, or self-attention layers, and with far fewer trainable parameters. With much fewer training parameters and higher representative power, our approach paves the way for broader adoption of INR models for generative modeling tasks in complex domains. The code is available at https://github.com/Rajhans0/Poly_INR
****

Neural Video Compression With Diverse Contexts
Jiahao Li, Bin Li, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22616-22626
For any video codecs, the coding efficiency highly relies on whether the current signal to be encoded can find the relevant contexts from the previous reconstructed signals. Traditional codec has verified more contexts bring substantial coding gain, but in a time-consuming manner. However, for the emerging neural video codec (NVC), its contexts are still limited, leading to low compression ratio. To boost NVC, this paper proposes increasing the context diversity in both temporal and spatial dimensions. First, we guide the model to learn hierarchical quality patterns across frames, which enriches long-term and yet high-quality temporal contexts. Furthermore, to tap the potential of optical flow-based coding framework, we introduce a group-based offset diversity where the cross-group interaction is proposed for better context mining. In addition, this paper also adopts a quadtree-based partition to increase spatial context diversity when encoding the latent representation in parallel. Experiments show that our codec obtains 23.5% bitrate saving over previous SOTA NVC. Better yet, our codec has surpassed the under-developing next generation traditional codec/ECM in both RGB and YUV420 colorspaces, in terms of PSNR. The codes are at https://github.com/microsoft/DCVC.
****

High-Frequency Stereo Matching Network
Haoliang Zhao, Huizhou Zhou, Yongjun Zhang, Jie Chen, Yitong Yang, Yong Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1327-1336
In the field of binocular stereo matching, remarkable progress has been made by iterative methods like RAFT-Stereo and CREStereo. However, most of these methods lose information during the iterative process, making it difficult to generate more detailed difference maps that take full advantage of high-frequency information. We propose the Decouple module to alleviate the problem of data coupling and allow features containing subtle details to transfer across the iterations which proves to alleviate the problem significantly in the ablations. To further capture high-frequency details, we propose a Normalization Refinement module that unifies the disparities as a proportion of the disparities over the width of the image, which address the problem of module failure in cross-domain scenarios. Further, with the above improvements, the ResNet-like feature extractor that has not been changed for years becomes a bottleneck. Towards this end, we proposed a multi-scale and multi-stage feature extractor that introduces the channel-wise self-attention mechanism which greatly addresses this bottleneck. Our method (DLNR) ranks 1st on the Middlebury leaderboard, significantly outperforming the next best method by 13.04%. Our method also achieves SOTA performance on the KITTI-2015 benchmark for D1-fg.
****

LayoutDM: Discrete Diffusion Model for Controllable Layout Generation
Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, Kota Yamaguchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10167-10176

Controllable layout generation aims at synthesizing plausible arrangement of element bounding boxes with optional constraints, such as type or position of a specific element. In this work, we try to solve a broad range of layout generation tasks in a single model that is based on discrete state-space diffusion models. Our model, named LayoutDM, naturally handles the structured layout data in the discrete representation and learns to progressively infer a noiseless layout from the initial input, where we model the layout corruption process by modality-wise discrete diffusion. For conditional generation, we propose to inject layout constraints in the form of masking or logit adjustment during inference. We show in the experiments that our LayoutDM successfully generates high-quality layouts and outperforms both task-specific and task-agnostic baselines on several layout tasks.
****
Markerless Camera-to-Robot Pose Estimation via Self-Supervised Sim-to-Real Transfer
Jingpei Lu, Florian Richter, Michael C. Yip; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21296-21306
Solving the camera-to-robot pose is a fundamental requirement for vision-based robot control, and is a process that takes considerable effort and cares to make accurate. Traditional approaches require modification of the robot via markers, and subsequent deep learning approaches enabled markerless feature extraction. Mainstream deep learning methods only use synthetic data and rely on Domain Randomization to fill the sim-to-real gap, because acquiring the 3D annotation is labor-intensive. In this work, we go beyond the limitation of 3D annotations for real-world data. We propose an end-to-end pose estimation framework that is capable of online camera-to-robot calibration and a self-supervised training method to scale the training to unlabeled real-world data. Our framework combines deep learning and geometric vision for solving the robot pose, and the pipeline is fully differentiable. To train the Camera-to-Robot Pose Estimation Network (CtRNet), we leverage foreground segmentation and differentiable rendering for image-level self-supervision. The pose prediction is visualized through a renderer and the image loss with the input image is back-propagated to train the neural network. Our experimental results on two public real datasets confirm the effectiveness of our approach over existing works. We also integrate our framework into a visual servoing system to demonstrate the promise of real-time precise robot pose estimation for automation tasks.
****
CARTO: Category and Joint Agnostic Reconstruction of ARTiculated Objects
Nick Heppert, Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Rares Andrei Ambrus, Jeannette Bohg, Abhinav Valada, Thomas Kollar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21201-21210
We present CARTO, a novel approach for reconstructing multiple articulated objects from a single stereo RGB observation. We use implicit object-centric representations and learn a single geometry and articulation decoder for multiple object categories. Despite training on multiple categories, our decoder achieves a comparable reconstruction accuracy to methods that train bespoke decoders separately for each category. Combined with our stereo image encoder we infer the 3D shape, 6D pose, size, joint type, and the joint state of multiple unknown objects in a single forward pass. Our method achieves a 20.4% absolute improvement in mAP 3D IOU50 for novel instances when compared to a two-stage pipeline. Inference time is fast and can run on a NVIDIA TITAN XP GPU at 1 HZ for eight or less objects present. While only trained on simulated data, CARTO transfers to real-world object instances. Code and evaluation data is available at: http://carto.cs.uni-freiburg.de
****
ShapeTalk: A Language Dataset and Framework for 3D Shape Edits and Deformations
Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, Leonidas Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12685-12694

Editing 3D geometry is a challenging task requiring specialized skills. In this work, we aim to facilitate the task of editing the geometry of 3D models through the use of natural language. For example, we may want to modify a 3D chair model to "make its legs thinner" or to "open a hole in its back". To tackle this problem in a manner that promotes open-ended language use and enables fine-grained shape edits, we introduce the most extensive existing corpus of natural language utterances describing shape differences: ShapeTalk. ShapeTalk contains over half a million discriminative utterances produced by contrasting the shapes of common 3D objects for a variety of object classes and degrees of similarity. We also introduce a generic framework, ChangeIt3D, which builds on ShapeTalk and can use an arbitrary 3D generative model of shapes to produce edits that align the output better with the edit or deformation description. Finally, we introduce metrics for the quantitative evaluation of language-assisted shape editing methods that reflect key desiderata within this editing setup. We note that ShapeTalk allows methods to be trained with explicit 3D-to-language data, bypassing the necessity of "lifting" 2D to 3D using methods like neural rendering, as required by extant 2D image-language foundation models. Our code and data are publicly available at https://changeit3d.github.io/.
****
Event-Guided Person Re-Identification via Sparse-Dense Complementary Learning
Chengzhi Cao, Xueyang Fu, Hongjian Liu, Yukun Huang, Kunyu Wang, Jiebo Luo, Zheng-Jun Zha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17990-17999
Video-based person re-identification (Re-ID) is a prominent computer vision topic due to its wide range of video surveillance applications. Most existing methods utilize spatial and temporal correlations in frame sequences to obtain discriminative person features. However, inevitable degradations, e.g., motion blur contained in frames often cause ambiguity texture noise and temporal disturbance, leading to the loss of identity-discriminating cues. Recently, a new bio-inspired sensor called event camera, which can asynchronously record intensity changes, brings new vitality to the Re-ID task. With the microsecond resolution and low latency, event cameras can accurately capture the movements of pedestrians even in the aforementioned degraded environments. Inspired by the properties of event cameras, in this work, we propose a Sparse-Dense Complementary Learning Framework, which effectively extracts identity features by fully exploiting the complementary information of dense frames and sparse events. Specifically, for frames, we build a CNN-based module to aggregate the dense features of pedestrian appearance step-by-step, while for event streams, we design a bio-inspired spiking neural backbone, which encodes event signals into sparse feature maps in a spiking form, to present the dynamic motion cues of pedestrians. Finally, a cross feature alignment module is constructed to complementarily fuse motion information from events and appearance cues from frames to enhance identity representation learning. Experiments on several benchmarks show that by employing events and SNN into Re-ID, our method significantly outperforms competitive methods.
****
Regularizing Second-Order Influences for Continual Learning
Zhicheng Sun, Yadong Mu, Gang Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20166-20175
Continual learning aims to learn on non-stationary data streams without catastrophically forgetting previous knowledge. Prevalent replay-based methods address this challenge by rehearsing on a small buffer holding the seen data, for which a delicate sample selection strategy is required. However, existing selection schemes typically seek only to maximize the utility of the ongoing selection, overlooking the interference between successive rounds of selection. Motivated by this, we dissect the interaction of sequential selection steps within a framework built on influence functions. We manage to identify a new class of second-order influences that will gradually amplify incidental bias in the replay buffer and compromise the selection process. To regularize the second-order effects, a novel selection objective is proposed, which also has clear connections to two widely adopted criteria. Furthermore, we present an efficient implementation for optim

izing the proposed criterion. Experiments on multiple continual learning benchmarks demonstrate the advantage of our approach over state-of-the-art methods. Code is available at https://github.com/feifeiobama/InfluenceCL.
****
Spatial-Then-Temporal Self-Supervised Learning for Video Correspondence
Rui Li, Dong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2279-2288
In low-level video analyses, effective representations are important to derive the correspondences between video frames. These representations have been learned in a self-supervised fashion from unlabeled images/videos, using carefully designed pretext tasks in some recent studies. However, the previous work concentrates on either spatial-discriminative features or temporal-repetitive features, with little attention to the synergy between spatial and temporal cues. To address this issue, we propose a novel spatial-then-temporal self-supervised learning method. Specifically, we firstly extract spatial features from unlabeled images via contrastive learning, and secondly enhance the features by exploiting the temporal cues in unlabeled videos via reconstructive learning. In the second step, we design a global correlation distillation loss to ensure the learning not to forget the spatial cues, and we design a local correlation distillation loss to combat the temporal discontinuity that harms the reconstruction. The proposed method outperforms the state-of-the-art self-supervised methods, as established by the experimental results on a series of correspondence-based video analysis tasks. Also, we performed ablation studies to verify the effectiveness of the two-step design as well as the distillation losses.
****
Super-Resolution Neural Operator
Min Wei, Xuesong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18247-18256
We propose Super-resolution Neural Operator (SRNO), a deep operator learning framework that can resolve high-resolution (HR) images at arbitrary scales from the low-resolution (LR) counterparts. Treating the LR-HR image pairs as continuous functions approximated with different grid sizes, SRNO learns the mapping between the corresponding function spaces. From the perspective of approximation theory, SRNO first embeds the LR input into a higher-dimensional latent representation space, trying to capture sufficient basis functions, and then iteratively approximates the implicit image function with a kernel integral mechanism, followed by a final dimensionality reduction step to generate the RGB representation at the target coordinates. The key characteristics distinguishing SRNO from prior continuous SR works are: 1) the kernel integral in each layer is efficiently implemented via the Galerkin-type attention, which possesses non-local properties in the spatial domain and therefore benefits the grid-free continuum; and 2) the multilayer attention architecture allows for the dynamic latent basis update, which is crucial for SR problems to "hallucinate" high-frequency information from the LR image. Experiments show that SRNO outperforms existing continuous SR methods in terms of both accuracy and running time. Our code is at https://github.com/2y7c3/Super-Resolution-Neural-Operator.
****
GradICON: Approximate Diffeomorphisms via Gradient Inverse Consistency
Lin Tian, Hastings Greer, François-Xavier Vialard, Roland Kwitt, Raúl San José Estépar, Richard Jarrett Rushmore, Nikolaos Makris, Sylvain Bouix, Marc Niethammer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18084-18094
We present an approach to learning regular spatial transformations between image pairs in the context of medical image registration. Contrary to optimization-based registration techniques and many modern learning-based methods, we do not directly penalize transformation irregularities but instead promote transformation regularity via an inverse consistency penalty. We use a neural network to predict a map between a source and a target image as well as the map when swapping the source and target images. Different from existing approaches, we compose these two resulting maps and regularize deviations of the Jacobian of this compositio

n from the identity matrix. This regularizer -- GradICON -- results in much better convergence when training registration models compared to promoting inverse consistency of the composition of maps directly while retaining the desirable implicit regularization effects of the latter. We achieve state-of-the-art registration performance on a variety of real-world medical image datasets using a single set of hyperparameters and a single non-dataset-specific training protocol. The code is available at https://github.com/uncbiag/ICON.
****
LP-DIF: Learning Local Pattern-Specific Deep Implicit Function for 3D Objects and Scenes
Meng Wang, Yu-Shen Liu, Yue Gao, Kanle Shi, Yi Fang, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21856-21865
Deep Implicit Function (DIF) has gained much popularity as an efficient 3D shape representation. To capture geometry details, current mainstream methods divide 3D shapes into local regions and then learn each one with a local latent code via a decoder, where the decoder shares the geometric similarities among different local regions. Although such local methods can capture more local details, a large diversity of different local regions increases the difficulty of learning an implicit function when treating all regions equally using only a single decoder. In addition, these local regions often exhibit imbalanced distributions, where certain regions have significantly fewer observations. This leads that fine geometry details could not be preserved well. To solve this problem, we propose a novel Local Pattern-specific Implicit Function, named LP-DIF, for representing a shape with some clusters of local regions and multiple decoders, where each decoder only focuses on one cluster of local regions which share a certain pattern. Specifically, we first extract local codes for all regions, and then cluster them into multiple groups in the latent space, where similar regions sharing a common pattern fall into one group. After that, we train multiple decoders for mining local patterns of different groups, which simplifies learning of fine geometric details by reducing the diversity of local regions seen by each decoder. To further alleviate the data-imbalance problem, we introduce a region re-weighting module to each pattern-specific decoder by kernel density estimator, which dynamically re-weights the regions during learning. Our LP-DIF can restore more geometry details, and thus improve the quality of 3D reconstruction. Experiments demonstrate that our method can achieve the state-of-the-art performance over previous methods. Code is available at https://github.com/gtyxyz/lpdif.
****
PeakConv: Learning Peak Receptive Field for Radar Semantic Segmentation
Liwen Zhang, Xinyan Zhang, Youcheng Zhang, Yufei Guo, Yuanpei Chen, Xuhui Huang, Zhe Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17577-17586
The modern machine learning-based technologies have shown considerable potential in automatic radar scene understanding. Among these efforts, radar semantic segmentation (RSS) can provide more refined and detailed information including the moving objects and background clutters within the effective receptive field of the radar. Motivated by the success of convolutional networks in various visual computing tasks, these networks have also been introduced to solve RSS task. However, neither the regular convolution operation nor the modified ones are specific to interpret radar signals. The receptive fields of existing convolutions are defined by the object presentation in optical signals, but these two signals have different perception mechanisms. In classic radar signal processing, the object signature is detected according to a local peak response, i.e., CFAR detection. Inspired by this idea, we redefine the receptive field of the convolution operation as the peak receptive field (PRF) and propose the peak convolution operation (PeakConv) to learn the object signatures in an end-to-end network. By incorporating the proposed PeakConv layers into the encoders, our RSS network can achieve better segmentation results compared with other SoTA methods on a multi-view real-measured dataset collected from an FMCW radar. Our code for PeakConv is available at https://github.com/zlw9161/PKC.

****
Unsupervised Contour Tracking of Live Cells by Mechanical and Cycle Consistency Losses
Junbong Jang, Kwonmoo Lee, Tae-Kyun Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 227-236
Analyzing the dynamic changes of cellular morphology is important for understanding the various functions and characteristics of live cells, including stem cells and metastatic cancer cells. To this end, we need to track all points on the highly deformable cellular contour in every frame of live cell video. Local shapes and textures on the contour are not evident, and their motions are complex, often with expansion and contraction of local contour features. The prior arts for optical flow or deep point set tracking are unsuited due to the fluidity of cells, and previous deep contour tracking does not consider point correspondence. We propose the first deep learning-based tracking of cellular (or more generally viscoelastic materials) contours with point correspondence by fusing dense representation between two contours with cross attention. Since it is impractical to manually label dense tracking points on the contour, unsupervised learning comprised of the mechanical and cyclical consistency losses is proposed to train our contour tracker. The mechanical loss forcing the points to move perpendicular to the contour effectively helps out. For quantitative evaluation, we labeled sparse tracking points along the contour of live cells from two live cell datasets taken with phase contrast and confocal fluorescence microscopes. Our contour tracker quantitatively outperforms compared methods and produces qualitatively more favorable results. Our code and data are publicly available at https://github.com/JunbongJang/contour-tracking/
****
Explaining Image Classifiers With Multiscale Directional Image Representation
Stefan Kolek, Robert Windesheim, Hector Andrade-Loarca, Gitta Kutyniok, Ron Levie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18600-18609
Image classifiers are known to be difficult to interpret and therefore require explanation methods to understand their decisions. We present ShearletX, a novel mask explanation method for image classifiers based on the shearlet transform -- a multiscale directional image representation. Current mask explanation methods are regularized by smoothness constraints that protect against undesirable fine-grained explanation artifacts. However, the smoothness of a mask limits its ability to separate fine-detail patterns, that are relevant for the classifier, from nearby nuisance patterns, that do not affect the classifier. ShearletX solves this problem by avoiding smoothness regularization all together, replacing it by shearlet sparsity constraints. The resulting explanations consist of a few edges, textures, and smooth parts of the original image, that are the most relevant for the decision of the classifier. To support our method, we propose a mathematical definition for explanation artifacts and an information theoretic score to evaluate the quality of mask explanations. We demonstrate the superiority of ShearletX over previous mask based explanation methods using these new metrics, and present exemplary situations where separating fine-detail patterns allows explaining phenomena that were not explainable before.
****
RGBD2: Generative Scene Synthesis via Incremental View Inpainting Using RGBD Diffusion Models
Jiabao Lei, Jiapeng Tang, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8422-8434
We address the challenge of recovering an underlying scene geometry and colors from a sparse set of RGBD view observations. In this work, we present a new solution termed RGBD2 that sequentially generates novel RGBD views along a camera trajectory, and the scene geometry is simply the fusion result of these views. More specifically, we maintain an intermediate surface mesh used for rendering new RGBD views, which subsequently becomes complete by an inpainting network; each rendered RGBD view is later back-projected as a partial surface and is supplemented into the intermediate mesh. The use of intermediate mesh and camera projection

 helps solve the tough problem of multi-view inconsistency. We practically imple
ment the RGBD inpainting network as a versatile RGBD diffusion model, which is p
reviously used for 2D generative modeling; we make a modification to its reverse
 diffusion process to enable our use. We evaluate our approach on the task of 3D
 scene synthesis from sparse RGBD inputs; extensive experiments on the ScanNet d
ataset demonstrate the superiority of our approach over existing ones. Project p
age: https://jblei.site/proj/rgbd-diffusion.
****
Distribution Shift Inversion for Out-of-Distribution Prediction
Runpeng Yu, Songhua Liu, Xingyi Yang, Xinchao Wang; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3592-360
2
Machine learning society has witnessed the emergence of a myriad of Out-of-Distr
ibution (OoD) algorithms, which address the distribution shift between the train
ing and the testing distribution by searching for a unified predictor or invaria
nt feature representation. However, the task of directly mitigating the distribu
tion shift in the unseen testing set is rarely investigated, due to the unavaila
bility of the testing distribution during the training phase and thus the imposs
ibility of training a distribution translator mapping between the training and t
esting distribution. In this paper, we explore how to bypass the requirement of
testing distribution for distribution translator training and make the distribut
ion translation useful for OoD prediction. We propose a portable Distribution Sh
ift Inversion (DSI) algorithm, in which, before being fed into the prediction mo
del, the OoD testing samples are first linearly combined with additional Gaussia
n noise and then transferred back towards the training distribution using a diff
usion model trained only on the source distribution. Theoretical analysis reveal
s the feasibility of our method. Experimental results, on both multiple-domain g
eneralization datasets and single-domain generalization datasets, show that our
method provides a general performance gain when plugged into a wide range of com
monly used OoD algorithms. Our code is available at https://github.com/yu-rp/Dis
tribution-Shift-Iverson https://github.com/yu-rp/Distribution-Shift-Iverson.
****
Deep Polarization Reconstruction With PDAVIS Events
Haiyang Mei, Zuowen Wang, Xin Yang, Xiaopeng Wei, Tobi Delbruck; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
 pp. 22149-22158
The polarization event camera PDAVIS is a novel bio-inspired neuromorphic vision
 sensor that reports both conventional polarization frames and asynchronous, con
tinuously per-pixel polarization brightness changes (polarization events) with f
ast temporal resolution and large dynamic range. A deep neural network method (P
olarization FireNet) was previously developed to reconstruct the polarization an
gle and degree from polarization events for bridging the gap between the polariz
ation event camera and mainstream computer vision. However, Polarization FireNet
 applies a network pre-trained for normal event-based frame reconstruction indep
endently on each of four channels of polarization events from four linear polari
zation angles, which ignores the correlations between channels and inevitably in
troduces content inconsistency between the four reconstructed frames, resulting
in unsatisfactory polarization reconstruction performance. In this work, we stri
ve to train an effective, yet efficient, DNN model that directly outputs polariz
ation from the input raw polarization events. To this end, we constructed the fi
rst large-scale event-to-polarization dataset, which we subsequently employed to
 train our events-to-polarization network E2P. E2P extracts rich polarization pa
tterns from input polarization events and enhances features through cross-modali
ty context integration. We demonstrate that E2P outperforms Polarization FireNet
 by a significant margin with no additional computing cost. Experimental results
 also show that E2P produces more accurate measurement of polarization than the
PDAVIS frames in challenging fast and high dynamic range scenes.
****
VideoTrack: Learning To Track Objects via Video Transformer
Fei Xie, Lei Chu, Jiahao Li, Yan Lu, Chao Ma; Proceedings of the IEEE/CVF Confer

ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22826-22835
Existing Siamese tracking methods, which are built on pair-wise matching between two single frames, heavily rely on additional sophisticated mechanism to exploit temporal information among successive video frames, hindering them from high efficiency and industrial deployments. In this work, we resort to sequence-level target matching that can encode temporal contexts into the spatial features through a neat feedforward video model. Specifically, we adapt the standard video transformer architecture to visual tracking by enabling spatiotemporal feature learning directly from frame-level patch sequences. To better adapt to the tracking task, we carefully blend the spatiotemporal information in the video clips through sequential multi-branch triplet blocks, which formulates a video transformer backbone. Our experimental study compares different model variants, such as tokenization strategies, hierarchical structures, and video attention schemes. Then, we propose a disentangled dual-template mechanism that decouples static and dynamic appearance changes over time, and reduces the temporal redundancy in video frames. Extensive experiments show that our method, named as VideoTrack, achieves state-of-the-art results while running in real-time.
****
System-Status-Aware Adaptive Network for Online Streaming Video Understanding
Lin Geng Foo, Jia Gong, Zhipeng Fan, Jun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10514-10523
Recent years have witnessed great progress in deep neural networks for real-time applications. However, most existing works do not explicitly consider the general case where the device's state and the available resources fluctuate over time, and none of them investigate or address the impact of varying computational resources for online video understanding tasks. This paper proposes a System-status-aware Adaptive Network (SAN) that considers the device's real-time state to provide high-quality predictions with low delay. Usage of our agent's policy improves efficiency and robustness to fluctuations of the system status. On two widely used video understanding tasks, SAN obtains state-of-the-art performance while constantly keeping processing delays low. Moreover, training such an agent on various types of hardware configurations is not easy as the labeled training data might not be available, or can be computationally prohibitive. To address this challenging problem, we propose a Meta Self-supervised Adaptation (MSA) method that adapts the agent's policy to new hardware configurations at test-time, allowing for easy deployment of the model onto other unseen hardware platforms.
****
Parallel Diffusion Models of Operator and Image for Blind Inverse Problems
Hyungjin Chung, Jeongsol Kim, Sehui Kim, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6059-6069
Diffusion model-based inverse problem solvers have demonstrated state-of-the-art performance in cases where the forward operator is known (i.e. non-blind). However, the applicability of the method to blind inverse problems has yet to be explored. In this work, we show that we can indeed solve a family of blind inverse problems by constructing another diffusion prior for the forward operator. Specifically, parallel reverse diffusion guided by gradients from the intermediate stages enables joint optimization of both the forward operator parameters as well as the image, such that both are jointly estimated at the end of the parallel reverse diffusion procedure. We show the efficacy of our method on two representative tasks --- blind deblurring, and imaging through turbulence --- and show that our method yields state-of-the-art performance, while also being flexible to be applicable to general blind inverse problems when we know the functional forms. Code available: https://github.com/BlindDPS/blind-dps
****
Local-Guided Global: Paired Similarity Representation for Visual Reinforcement Learning
Hyesong Choi, Hunsang Lee, Wonil Song, Sangryul Jeon, Kwanghoon Sohn, Dongbo Min; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15072-15082

Recent vision-based reinforcement learning (RL) methods have found extracting hi
gh-level features from raw pixels with self-supervised learning to be effective
in learning policies. However, these methods focus on learning global representa
tions of images, and disregard local spatial structures present in the consecuti
vely stacked frames. In this paper, we propose a novel approach, termed self-sup
ervised Paired Similarity Representation Learning (PSRL) for effectively encodin
g spatial structures in an unsupervised manner. Given the input frames, the late
nt volumes are first generated individually using an encoder, and they are used
to capture the variance in terms of local spatial structures, i.e., corresponden
ce maps among multiple frames. This enables for providing plenty of fine-grained
samples for training the encoder of deep RL. We further attempt to learn the gl
obal semantic representations in the global prediction module that predicts futu
re state representations using action vector as a medium. The proposed method im
poses similarity constraints on the three latent volumes; transformed query repr
esentations by estimated pixel-wise correspondence, predicted query representati
ons from the global prediction model, and target representations of future state
, guiding global prediction with locality-inherent volume. Experimental results
on complex tasks in Atari Games and DeepMind Control Suite demonstrate that the
RL methods are significantly boosted by the proposed self-supervised learning of
structured representations.
****
Semidefinite Relaxations for Robust Multiview Triangulation
Linus Härenstam-Nielsen, Niclas Zeller, Daniel Cremers; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 749-
757
We propose an approach based on convex relaxations for certifiably optimal robus
t multiview triangulation. To this end, we extend existing relaxation approaches
to non-robust multiview triangulation by incorporating a least squares cost fun
ction. We propose two formulations, one based on epipolar constraints and one ba
sed on fractional reprojection constraints. The first is lower dimensional and r
emains tight under moderate noise and outlier levels, while the second is higher
dimensional and therefore slower but remains tight even under extreme noise and
outlier levels. We demonstrate through extensive experiments that the proposed
approaches allow us to compute provably optimal reconstructions even under signi
ficant noise and a large percentage of outliers.
****
Distilling Self-Supervised Vision Transformers for Weakly-Supervised Few-Shot Cl
assification & Segmentation
Dahyun Kang, Piotr Koniusz, Minsu Cho, Naila Murray; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19627-1
9638
We address the task of weakly-supervised few-shot image classification and segme
ntation, by leveraging a Vision Transformer (ViT) pretrained with self-supervisi
on. Our proposed method takes token representations from the self-supervised ViT
and leverages their correlations, via self-attention, to produce classification
and segmentation predictions through separate task heads. Our model is able to
effectively learn to perform classification and segmentation in the absence of p
ixel-level labels during training, using only image-level labels. To do this it
uses attention maps, created from tokens generated by the self-supervised ViT ba
ckbone, as pixel-level pseudo-labels. We also explore a practical setup with "mi
xed" supervision, where a small number of training images contains ground-truth
pixel-level labels and the remaining images have only image-level labels. For th
is mixed setup, we propose to improve the pseudo-labels using a pseudo-label enh
ancer that was trained using the available ground-truth pixel-level labels. Expe
riments on Pascal-5i and COCO-20i demonstrate significant performance gains in a
variety of supervision settings, and in particular when little-to-no pixel-leve
l labels are available.
****
FFCV: Accelerating Training by Removing Data Bottlenecks
Guillaume Leclerc, Andrew Ilyas, Logan Engstrom, Sung Min Park, Hadi Salman, Ale

ksander M■dry; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2023, pp. 12011-12020
We present FFCV, a library for easy, fast, resource-efficient training of machin
e learning models. FFCV speeds up model training by eliminating (often subtle) d
ata bottlenecks from the training process. In particular, we combine techniques
such as an efficient file storage format, caching, data pre-loading, asynchronou
s data transfer, and just-in-time compilation to (a) make data loading and trans
fer significantly more efficient, ensuring that GPUs can reach full utilization;
and (b) offload as much data processing as possible to the CPU asynchronously,
freeing GPU up capacity for training. Using FFCV, we train ResNet-18 and ResNet-
50 on the ImageNet dataset with a state-of-the-art tradeoff between accuracy and
training time. For example, across the range of ResNet-50 models we test, we ob
tain the same accuracy as the best baselines in half the time. We demonstrate FF
CV's performance, ease-of-use, extensibility, and ability to adapt to resource c
onstraints through several case studies.
****
Collaborative Noisy Label Cleaner: Learning Scene-Aware Trailers for Multi-Modal
Highlight Detection in Movies
Bei Gan, Xiujun Shu, Ruizhi Qiao, Haoqian Wu, Keyu Chen, Hanjun Li, Bo Ren; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2023, pp. 18898-18907
Movie highlights stand out of the screenplay for efficient browsing and play a c
rucial role on social media platforms. Based on existing efforts, this work has
two observations: (1) For different annotators, labeling highlight has uncertain
ty, which leads to inaccurate and time-consuming annotations. (2) Besides previo
us supervised or unsupervised settings, some existing video corpora can be usefu
l, e.g., trailers, but they are often noisy and incomplete to cover the full hig
hlights. In this work, we study a more practical and promising setting, i.e., re
formulating highlight detection as "learning with noisy labels". This setting do
es not require time-consuming manual annotations and can fully utilize existing
abundant video corpora. First, based on movie trailers, we leverage scene segmen
tation to obtain complete shots, which are regarded as noisy labels. Then, we pr
opose a Collaborative noisy Label Cleaner (CLC) framework to learn from noisy hi
ghlight moments. CLC consists of two modules: augmented cross-propagation (ACP)
and multi-modality cleaning (MMC). The former aims to exploit the closely relate
d audio-visual signals and fuse them to learn unified multi-modal representation
s. The latter aims to achieve cleaner highlight labels by observing the changes
in losses among different modalities. To verify the effectiveness of CLC, we fur
ther collect a large-scale highlight dataset named MovieLights. Comprehensive ex
periments on MovieLights and YouTube Highlights datasets demonstrate the effecti
veness of our approach. Code has been made available at: https://github.com/Tenc
entYoutuResearch/HighlightDetection-CLC
****
Modeling Video As Stochastic Processes for Fine-Grained Video Representation Lea
rning
Heng Zhang, Daqing Liu, Qi Zheng, Bing Su; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2225-2234
A meaningful video is semantically coherent and changes smoothly. However, most
existing fine-grained video representation learning methods learn frame-wise fea
tures by aligning frames across videos or exploring relevance between multiple v
iews, neglecting the inherent dynamic process of each video. In this paper, we p
ropose to learn video representations by modeling Video as Stochastic Processes
(VSP) via a novel process-based contrastive learning framework, which aims to di
scriminate between video processes and simultaneously capture the temporal dynam
ics in the processes. Specifically, we enforce the embeddings of the frame seque
nce of interest to approximate a goal-oriented stochastic process, i.e., Brownia
n bridge, in the latent space via a process-based contrastive loss. To construct
the Brownian bridge, we adapt specialized sampling strategies under different a
nnotations for both self-supervised and weakly-supervised learning. Experimental
results on four datasets show that VSP stands as a state-of-the-art method for

various video understanding tasks, including phase progression, phase classification and frame retrieval. Code is available at 'https://github.com/hengRUC/VSP'.
****
ContraNeRF: Generalizable Neural Radiance Fields for Synthetic-to-Real Novel View Synthesis via Contrastive Learning
Hao Yang, Lanqing Hong, Aoxue Li, Tianyang Hu, Zhenguo Li, Gim Hee Lee, Liwei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16508-16517
Although many recent works have investigated generalizable NeRF-based novel view synthesis for unseen scenes, they seldom consider the synthetic-to-real generalization, which is desired in many practical applications. In this work, we first investigate the effects of synthetic data in synthetic-to-real novel view synthesis and surprisingly observe that models trained with synthetic data tend to produce sharper but less accurate volume densities. For pixels where the volume densities are correct, fine-grained details will be obtained. Otherwise, severe artifacts will be produced. To maintain the advantages of using synthetic data while avoiding its negative effects, we propose to introduce geometry-aware contrastive learning to learn multi-view consistent features with geometric constraints. Meanwhile, we adopt cross-view attention to further enhance the geometry perception of features by querying features across input views. Experiments demonstrate that under the synthetic-to-real setting, our method can render images with higher quality and better fine-grained details, outperforming existing generalizable novel view synthesis methods in terms of PSNR, SSIM, and LPIPS. When trained on real data, our method also achieves state-of-the-art results. https://haoy945.github.io/contranerf/
****
Region-Aware Pretraining for Open-Vocabulary Object Detection With Vision Transformers
Dahun Kim, Anelia Angelova, Weicheng Kuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11144-11154
We present Region-aware Open-vocabulary Vision Transformers (RO-ViT) -- a contrastive image-text pretraining recipe to bridge the gap between image-level pretraining and open-vocabulary object detection. At the pretraining phase, we propose to randomly crop and resize regions of positional embeddings instead of using the whole image positional embeddings. This better matches the use of positional embeddings at region-level in the detection finetuning phase. In addition, we replace the common softmax cross entropy loss in contrastive learning with focal loss to better learn the informative yet difficult examples. Finally, we leverage recent advances in novel object proposals to improve open-vocabulary detection finetuning. We evaluate our full model on the LVIS and COCO open-vocabulary detection benchmarks and zero-shot transfer. RO-ViT achieves a state-of-the-art 32.1 APr on LVIS, surpassing the best existing approach by +5.8 points in addition to competitive zero-shot transfer detection. Surprisingly, RO-ViT improves the image-level representation as well and achieves the state of the art on 9 out of 12 metrics on COCO and Flickr image-text retrieval benchmarks, outperforming competitive approaches with larger models.
****
PaletteNeRF: Palette-Based Appearance Editing of Neural Radiance Fields
Zhengfei Kuang, Fujun Luan, Sai Bi, Zhixin Shu, Gordon Wetzstein, Kalyan Sunkavalli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20691-20700
Recent advances in neural radiance fields have enabled the high-fidelity 3D reconstruction of complex scenes for novel view synthesis. However, it remains underexplored how the appearance of such representations can be efficiently edited while maintaining photorealism. In this work, we present PaletteNeRF, a novel method for photorealistic appearance editing of neural radiance fields (NeRF) based on 3D color decomposition. Our method decomposes the appearance of each 3D point into a linear combination of palette-based bases (i.e., 3D segmentations defined by a group of NeRF-type functions) that are shared across the scene. While our palette-based bases are view-independent, we also predict a view-dependent func

tion to capture the color residual (e.g., specular shading). During training, we jointly optimize the basis functions and the color palettes, and we also introduce novel regularizers to encourage the spatial coherence of the decomposition. Our method allows users to efficiently edit the appearance of the 3D scene by modifying the color palettes. We also extend our framework with compressed semantic features for semantic-aware appearance editing. We demonstrate that our technique is superior to baseline methods both quantitatively and qualitatively for appearance editing of complex real-world scenes.
****

Towards Unsupervised Object Detection From LiDAR Point Clouds
Lunjun Zhang, Anqi Joyce Yang, Yuwen Xiong, Sergio Casas, Bin Yang, Mengye Ren, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9317-9328
In this paper, we study the problem of unsupervised object detection from 3D point clouds in self-driving scenes. We present a simple yet effective method that exploits (i) point clustering in near-range areas where the point clouds are dense, (ii) temporal consistency to filter out noisy unsupervised detections, (iii) translation equivariance of CNNs to extend the auto-labels to long range, and (iv) self-supervision for improving on its own. Our approach, OYSTER (Object Discovery via Spatio-Temporal Refinement), does not impose constraints on data collection (such as repeated traversals of the same location), is able to detect objects in a zero-shot manner without supervised finetuning (even in sparse, distant regions), and continues to self-improve given more rounds of iterative self-training. To better measure model performance in self-driving scenarios, we propose a new planning-centric perception metric based on distance-to-collision. We demonstrate that our unsupervised object detector significantly outperforms unsupervised baselines on PandaSet and Argoverse 2 Sensor dataset, showing promise that self-supervision combined with object priors can enable object discovery in the wild. For more information, visit the project website: https://waabi.ai/research/oyster.
****

Contrastive Mean Teacher for Domain Adaptive Object Detectors
Shengcao Cao, Dhiraj Joshi, Liang-Yan Gui, Yu-Xiong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23839-23848
Object detectors often suffer from the domain gap between training (source domain) and real-world applications (target domain). Mean-teacher self-training is a powerful paradigm in unsupervised domain adaptation for object detection, but it struggles with low-quality pseudo-labels. In this work, we identify the intriguing alignment and synergy between mean-teacher self-training and contrastive learning. Motivated by this, we propose Contrastive Mean Teacher (CMT) -- a unified, general-purpose framework with the two paradigms naturally integrated to maximize beneficial learning signals. Instead of using pseudo-labels solely for final predictions, our strategy extracts object-level features using pseudo-labels and optimizes them via contrastive learning, without requiring labels in the target domain. When combined with recent mean-teacher self-training methods, CMT leads to new state-of-the-art target-domain performance: 51.9% mAP on Foggy Cityscapes, outperforming the previously best by 2.1% mAP. Notably, CMT can stabilize performance and provide more significant gains as pseudo-label noise increases.
****

Learning Transferable Spatiotemporal Representations From Natural Script Knowledge
Ziyun Zeng, Yuying Ge, Xihui Liu, Bin Chen, Ping Luo, Shu-Tao Xia, Yixiao Ge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23079-23089
Pre-training on large-scale video data has become a common recipe for learning transferable spatiotemporal representations in recent years. Despite some progress, existing methods are mostly limited to highly curated datasets (e.g., K400) and exhibit unsatisfactory out-of-the-box representations. We argue that it is due to the fact that they only capture pixel-level knowledge rather than spatiotem

poral semantics, which hinders further progress in video understanding. Inspired by the great success of image-text pre-training (e.g., CLIP), we take the first step to exploit language semantics to boost transferable spatiotemporal representation learning. We introduce a new pretext task, Turning to Video for Transcript Sorting (TVTS), which sorts shuffled ASR scripts by attending to learned video representations. We do not rely on descriptive captions and learn purely from video, i.e., leveraging the natural transcribed speech knowledge to provide noisy but useful semantics over time. Our method enforces the vision model to contextualize what is happening over time so that it can re-organize the narrative transcripts, and can seamlessly apply to large-scale uncurated video data in the real world. Our method demonstrates strong out-of-the-box spatiotemporal representations on diverse benchmarks, e.g., +13.6% gains over VideoMAE on SSV2 via linear probing. The code is available at https://github.com/TencentARC/TVTS.
****
NeRF-DS: Neural Radiance Fields for Dynamic Specular Objects
Zhiwen Yan, Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8285-8295
Dynamic Neural Radiance Field (NeRF) is a powerful algorithm capable of rendering photo-realistic novel view images from a monocular RGB video of a dynamic scene. Although it warps moving points across frames from the observation spaces to a common canonical space for rendering, dynamic NeRF does not model the change of the reflected color during the warping. As a result, this approach often fails drastically on challenging specular objects in motion. We address this limitation by reformulating the neural radiance field function to be conditioned on surface position and orientation in the observation space. This allows the specular surface at different poses to keep the different reflected colors when mapped to the common canonical space. Additionally, we add the mask of moving objects to guide the deformation field. As the specular surface changes color during motion, the mask mitigates the problem of failure to find temporal correspondences with only RGB supervision. We evaluate our model based on the novel view synthesis quality with a self-collected dataset of different moving specular objects in realistic environments. The experimental results demonstrate that our method significantly improves the reconstruction quality of moving specular objects from monocular RGB videos compared to the existing NeRF models. Our code and data are available at the project website https://github.com/JokerYan/NeRF-DS.
****
M6Doc: A Large-Scale Multi-Format, Multi-Type, Multi-Layout, Multi-Language, Multi-Annotation Category Dataset for Modern Document Layout Analysis
Hiuyi Cheng, Peirong Zhang, Sihang Wu, Jiaxin Zhang, Qiyuan Zhu, Zecheng Xie, Jing Li, Kai Ding, Lianwen Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15138-15147
Document layout analysis is a crucial prerequisite for document understanding, including document retrieval and conversion. Most public datasets currently contain only PDF documents and lack realistic documents. Models trained on these datasets may not generalize well to real-world scenarios. Therefore, this paper introduces a large and diverse document layout analysis dataset called M^6-Doc. The M^6 designation represents six properties: (1) Multi-Format (including scanned, photographed, and PDF documents); (2) Multi-Type (such as scientific articles, textbooks, books, test papers, magazines, newspapers, and notes); (3) Multi-Layout (rectangular, Manhattan, non-Manhattan, and multi-column Manhattan); (4) Multi-Language (Chinese and English); (5) Multi-Annotation Category (74 types of annotation labels with 237,116 annotation instances in 9,080 manually annotated pages); and (6) Modern documents. Additionally, we propose a transformer-based document layout analysis method called TransDLANet, which leverages an adaptive element matching mechanism that enables query embedding to better match ground truth to improve recall, and constructs a segmentation branch for more precise document image instance segmentation. We conduct a comprehensive evaluation of M^6-Doc with various layout analysis methods and demonstrate its effectiveness. TransDLANet achieves state-of-the-art performance on M^6-Doc with 64.5% mAP. The M^6-Doc dataset will be available at https://github.com/HCIILAB/M6Doc.

****
RealFusion: 360deg Reconstruction of Any Object From a Single Image
Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8446-8455
We consider the problem of reconstructing a full 360deg photographic model of an object from a single image of it. We do so by fitting a neural radiance field to the image, but find this problem to be severely ill-posed. We thus take an off-the-self conditional image generator based on diffusion and engineer a prompt that encourages it to "dream up" novel views of the object. Using the recent DreamFusion method, we fuse the given input view, the conditional prior, and other regularizers in a final, consistent reconstruction. We demonstrate state-of-the-art reconstruction results on benchmark images when compared to prior methods for monocular 3D reconstruction of objects. Qualitatively, our reconstructions provide a faithful match of the input view and a plausible extrapolation of its appearance and 3D shape, including to the side of the object not visible in the image.
****
CiCo: Domain-Aware Sign Language Retrieval via Cross-Lingual Contrastive Learning
Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, Wenqiang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19016-19026
This work focuses on sign language retrieval--a recently proposed task for sign language understanding. Sign language retrieval consists of two sub-tasks: text-to-sign-video (T2V) retrieval and sign-video-to-text (V2T) retrieval. Different from traditional video-text retrieval, sign language videos, not only contain visual signals but also carry abundant semantic meanings by themselves due to the fact that sign languages are also natural languages. Considering this character, we formulate sign language retrieval as a cross-lingual retrieval problem as well as a video-text retrieval task. Concretely, we take into account the linguistic properties of both sign languages and natural languages, and simultaneously identify the fine-grained cross-lingual (i.e., sign-to-word) mappings while contrasting the texts and the sign videos in a joint embedding space. This process is termed as cross-lingual contrastive learning. Another challenge is raised by the data scarcity issue--sign language datasets are orders of magnitude smaller in scale than that of speech recognition. We alleviate this issue by adopting a domain-agnostic sign encoder pre-trained on large-scale sign videos into the target domain via pseudo-labeling. Our framework, termed as domain-aware sign language retrieval via Cross-lingual Contrastive learning or CiCo for short, outperforms the pioneering method by large margins on various datasets, e.g., +22.4 T2V and +28.0 V2T R@1 improvements on How2Sign dataset, and +13.7 T2V and +17.1 V2T R@1 improvements on PHOENIX-2014T dataset. Code and models are available at: https://github.com/FangyunWei/SLRT.
****
Relational Space-Time Query in Long-Form Videos
Xitong Yang, Fu-Jen Chu, Matt Feiszli, Raghav Goyal, Lorenzo Torresani, Du Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6398-6408
Egocentric videos are often available in the form of uninterrupted, uncurated long videos capturing the camera wearers' daily life activities.Understanding these videos requires models to be able to reason about activities, objects, and their interactions. However, current video benchmarks study these problems independently and under short, curated clips. In contrast, real-world applications, e.g., AR assistants, require bundling these problems for both model development and evaluation. In this paper, we propose to study these problems in a joint framework for long video understanding. Our contributions are three-fold. First, we propose an integrated framework, namely Relational Space-Time Query (ReST), for evaluating video understanding models via templated spatiotemporal queries. Second, we introduce two new benchmarks, ReST-ADL and ReST-Ego4D, which augment the exi

sting egocentric video datasets with abundant query annotations generated by the ReST framework. Finally, we present a set of baselines and in-depth analysis on the two benchmarks and provide insights about the query tasks. We view our integrated framework and benchmarks as a step towards comprehensive, multi-step reasoning in long videos, and believe it will facilitate the development of next generations of video understanding models.
****
LargeKernel3D: Scaling Up Kernels in 3D Sparse CNNs
Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13488-13498
Recent advance in 2D CNNs has revealed that large kernels are important. However, when directly applying large convolutional kernels in 3D CNNs, severe difficulties are met, where those successful module designs in 2D become surprisingly ineffective on 3D networks, including the popular depth-wise convolution. To address this vital challenge, we instead propose the spatial-wise partition convolution and its large-kernel module. As a result, it avoids the optimization and efficiency issues of naive 3D large kernels. Our large-kernel 3D CNN network, LargeKernel3D, yields notable improvement in 3D tasks of semantic segmentation and object detection. It achieves 73.9% mIoU on the ScanNetv2 semantic segmentation and 72.8% NDS nuScenes object detection benchmarks, ranking 1st on the nuScenes LIDAR leaderboard. The performance further boosts to 74.2% NDS with a simple multi-modal fusion. In addition, LargeKernel3D can be scaled to 17x17x17 kernel size on Waymo 3D object detection. For the first time, we show that large kernels are feasible and essential for 3D visual tasks.
****
Video Dehazing via a Multi-Range Temporal Alignment Network With Physical Prior
Jiaqi Xu, Xiaowei Hu, Lei Zhu, Qi Dou, Jifeng Dai, Yu Qiao, Pheng-Ann Heng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18053-18062
Video dehazing aims to recover haze-free frames with high visibility and contrast. This paper presents a novel framework to effectively explore the physical haze priors and aggregate temporal information. Specifically, we design a memory-based physical prior guidance module to encode the prior-related features into long-range memory. Besides, we formulate a multi-range scene radiance recovery module to capture space-time dependencies in multiple space-time ranges, which helps to effectively aggregate temporal information from adjacent frames. Moreover, we construct the first large-scale outdoor video dehazing benchmark dataset, which contains videos in various real-world scenarios. Experimental results on both synthetic and real conditions show the superiority of our proposed method.
****
3D Concept Learning and Reasoning From Multi-View Images
Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B. Tenenbaum, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9202-9212
Humans are able to accurately reason in 3D by gathering multi-view observations of the surrounding world. Inspired by this insight, we introduce a new large-scale benchmark for 3D multi-view visual question answering (3DMV-VQA). This dataset is collected by an embodied agent actively moving and capturing RGB images in an environment using the Habitat simulator. In total, it consists of approximately 5k scenes, 600k images, paired with 50k questions. We evaluate various state-of-the-art models for visual reasoning on our benchmark and find that they all perform poorly. We suggest that a principled approach for 3D reasoning from multi-view images should be to infer a compact 3D representation of the world from the multi-view images, which is further grounded on open-vocabulary semantic concepts, and then to execute reasoning on these 3D representations. As the first step towards this approach, we propose a novel 3D concept learning and reasoning (3D-CLR) framework that seamlessly combines these components via neural fields, 2D pre-trained vision-language models, and neural reasoning operators. Experimental results suggest that our framework outperforms baseline models by a large marg

in, but the challenge remains largely unsolved. We further perform an in-depth a nalysis of the challenges and highlight potential future directions.
****
BiFormer: Learning Bilateral Motion Estimation via Bilateral Transformer for 4K Video Frame Interpolation
Junheum Park, Jintae Kim, Chang-Su Kim; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1568-1577
A novel 4K video frame interpolator based on bilateral transformer (BiFormer) is proposed in this paper, which performs three steps: global motion estimation, l ocal motion refinement, and frame synthesis. First, in global motion estimation, we predict symmetric bilateral motion fields at a coarse scale. To this end, we propose BiFormer, the first transformer-based bilateral motion estimator. Secon d, we refine the global motion fields efficiently using blockwise bilateral cost volumes (BBCVs). Third, we warp the input frames using the refined motion field s and blend them to synthesize an intermediate frame. Extensive experiments demo nstrate that the proposed BiFormer algorithm achieves excellent interpolation pe rformance on 4K datasets. The source codes are available at https://github.com/J unHeum/BiFormer.
****
Integrally Pre-Trained Transformer Pyramid Networks
Yunjie Tian, Lingxi Xie, Zhaozhi Wang, Longhui Wei, Xiaopeng Zhang, Jianbin Jiao , Yaowei Wang, Qi Tian, Qixiang Ye; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2023, pp. 18610-18620
In this paper, we present an integral pre-training framework based on masked ima ge modeling (MIM). We advocate for pre-training the backbone and neck jointly so that the transfer gap between MIM and downstream recognition tasks is minimal. We make two technical contributions. First, we unify the reconstruction and reco gnition necks by inserting a feature pyramid into the pre-training stage. Second , we complement mask image modeling (MIM) with masked feature modeling (MFM) tha t offers multi-stage supervision to the feature pyramid. The pre-trained models, termed integrally pre-trained transformer pyramid networks (iTPNs), serve as po werful foundation models for visual recognition. In particular, the base/large-l evel iTPN achieves an 86.2%/87.8% top-1 accuracy on ImageNet-1K, a 53.2%/55.6% b ox AP on COCO object detection with 1x training schedule using Mask-RCNN, and a 54.7%/57.7% mIoU on ADE20K semantic segmentation using UPerHead -- all these res ults set new records. Our work inspires the community to work on unifying upstre am pre-training and downstream fine-tuning tasks. Code is available at https://g ithub.com/sunsmarterjie/iTPN.
****
Soft Augmentation for Image Classification
Yang Liu, Shen Yan, Laura Leal-Taixé, James Hays, Deva Ramanan; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16241-16250
Modern neural networks are over-parameterized and thus rely on strong regulariza tion such as data augmentation and weight decay to reduce overfitting and improv e generalization. The dominant form of data augmentation applies invariant trans forms, where the learning target of a sample is invariant to the transform appli ed to that sample. We draw inspiration from human visual classification studies and propose generalizing augmentation with invariant transforms to soft augmenta tion where the learning target softens non-linearly as a function of the degree of the transform applied to the sample: e.g., more aggressive image crop augment ations produce less confident learning targets. We demonstrate that soft targets allow for more aggressive data augmentation, offer more robust performance boos ts, work with other augmentation policies, and interestingly, produce better cal ibrated models (since they are trained to be less confident on aggressively crop ped/occluded examples). Combined with existing aggressive augmentation strategie s, soft targets 1) double the top-1 accuracy boost across Cifar-10, Cifar-100, I mageNet-1K, and ImageNet-V2, 2) improve model occlusion performance by up to 4x, and 3) half the expected calibration error (ECE). Finally, we show that soft au gmentation generalizes to self-supervised classification tasks.

****
Learning From Unique Perspectives: User-Aware Saliency Modeling
Shi Chen, Nachiappan Valliappan, Shaolei Shen, Xinyu Ye, Kai Kohlhoff, Junfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2701-2710
Everyone is unique. Given the same visual stimuli, people's attention is driven by both salient visual cues and their own inherent preferences. Knowledge of visual preferences not only facilitates understanding of fine-grained attention patterns of diverse users, but also has the potential of benefiting the development of customized applications. Nevertheless, existing saliency models typically limit their scope to attention as it applies to the general population and ignore the variability between users' behaviors. In this paper, we identify the critical roles of visual preferences in attention modeling, and for the first time study the problem of user-aware saliency modeling. Our work aims to advance attention research from three distinct perspectives: (1) We present a new model with the flexibility to capture attention patterns of various combinations of users, so that we can adaptively predict personalized attention, user group attention, and general saliency at the same time with one single model; (2) To augment models with knowledge about the composition of attention from different users, we further propose a principled learning method to understand visual attention in a progressive manner; and (3) We carry out extensive analyses on publicly available saliency datasets to shed light on the roles of visual preferences. Experimental results on diverse stimuli, including naturalistic images and web pages, demonstrate the advantages of our method in capturing the distinct visual behaviors of different users and the general saliency of visual stimuli.
****
PREIM3D: 3D Consistent Precise Image Attribute Editing From a Single Image
Jianhui Li, Jianmin Li, Haoji Zhang, Shilong Liu, Zhengyi Wang, Zihao Xiao, Kaiwen Zheng, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8549-8558
We study the 3D-aware image attribute editing problem in this paper, which has wide applications in practice. Recent methods solved the problem by training a shared encoder to map images into a 3D generator's latent space or by per-image latent code optimization and then edited images in the latent space. Despite their promising results near the input view, they still suffer from the 3D inconsistency of produced images at large camera poses and imprecise image attribute editing, like affecting unspecified attributes during editing. For more efficient image inversion, we train a shared encoder for all images. To alleviate 3D inconsistency at large camera poses, we propose two novel methods, an alternating training scheme and a multi-view identity loss, to maintain 3D consistency and subject identity. As for imprecise image editing, we attribute the problem to the gap between the latent space of real images and that of generated images. We compare the latent space and inversion manifold of GAN models and demonstrate that editing in the inversion manifold can achieve better results in both quantitative and qualitative evaluations. Extensive experiments show that our method produces more 3D consistent images and achieves more precise image editing than previous work. Source code and pretrained models can be found on our project page: https://mybabyyh.github.io/Preim3D.
****
MaskSketch: Unpaired Structure-Guided Masked Image Generation
Dina Bashkirova, José Lezama, Kihyuk Sohn, Kate Saenko, Irfan Essa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1879-1889
Recent conditional image generation methods produce images of remarkable diversity, fidelity and realism. However, the majority of these methods allow conditioning only on labels or text prompts, which limits their level of control over the generation result. In this paper, we introduce MaskSketch, an image generation method that allows spatial conditioning of the generation result using a guiding sketch as an extra conditioning signal during sampling. MaskSketch utilizes a pre-trained masked generative transformer, requiring no model training or paired

supervision, and works with input sketches of different levels of abstraction. We show that intermediate self-attention maps of a masked generative transformer encode important structural information of the input image, such as scene layout and object shape, and we propose a novel sampling method based on this observation to enable structure-guided generation. Our results show that MaskSketch achieves high image realism and fidelity to the guiding structure. Evaluated on standard benchmark datasets, MaskSketch outperforms state-of-the-art methods for sketch-to-image translation, as well as unpaired image-to-image translation approaches. The code can be found on our project website: https://masksketch.github.io/
****
Open-Vocabulary Point-Cloud Object Detection Without 3D Annotation
Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1190-1199
The goal of open-vocabulary detection is to identify novel objects based on arbitrary textual descriptions. In this paper, we address open-vocabulary 3D point-cloud detection by a dividing-and-conquering strategy, which involves: 1) developing a point-cloud detector that can learn a general representation for localizing various objects, and 2) connecting textual and point-cloud representations to enable the detector to classify novel object categories based on text prompting. Specifically, we resort to rich image pre-trained models, by which the point-cloud detector learns localizing objects under the supervision of predicted 2D bounding boxes from 2D pre-trained detectors. Moreover, we propose a novel de-biased triplet cross-modal contrastive learning to connect the modalities of image, point-cloud and text, thereby enabling the point-cloud detector to benefit from vision-language pre-trained models, i.e., CLIP. The novel use of image and vision-language pre-trained models for point-cloud detectors allows for open-vocabulary 3D object detection without the need for 3D annotations. Experiments demonstrate that the proposed method improves at least 3.03 points and 7.47 points over a wide range of baselines on the ScanNet and SUN RGB-D datasets, respectively. Furthermore, we provide a comprehensive analysis to explain why our approach works.
****
Adaptive Channel Sparsity for Federated Learning Under System Heterogeneity
Dongping Liao, Xitong Gao, Yiren Zhao, Cheng-Zhong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20432-20441
Owing to the non-i.i.d. nature of client data, channel neurons in federated-learned models may specialize to distinct features for different clients. Yet, existing channel-sparse federated learning (FL) algorithms prescribe fixed sparsity strategies for client models, and may thus prevent clients from training channel neurons collaboratively. To minimize the impact of sparsity on FL convergence, we propose Flado to improve the alignment of client model update trajectories by tailoring the sparsities of individual neurons in each client. Empirical results show that while other sparse methods are surprisingly impactful to convergence, Flado can not only attain the highest task accuracies with unlimited budget across a range of datasets, but also significantly reduce the amount of FLOPs required for training more than by 10x under the same communications budget, and push the Pareto frontier of communication/computation trade-off notably further than competing FL algorithms.
****
Detecting Backdoors in Pre-Trained Encoders
Shiwei Feng, Guanhong Tao, Siyuan Cheng, Guangyu Shen, Xiangzhe Xu, Yingqi Liu, Kaiyuan Zhang, Shiqing Ma, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16352-16362
Self-supervised learning in computer vision trains on unlabeled data, such as images or (image, text) pairs, to obtain an image encoder that learns high-quality embeddings for input data. Emerging backdoor attacks towards encoders expose crucial vulnerabilities of self-supervised learning, since downstream classifiers (even further trained on clean data) may inherit backdoor behaviors from encoder

s. Existing backdoor detection methods mainly focus on supervised learning settings and cannot handle pre-trained encoders especially when input labels are not available. In this paper, we propose DECREE, the first backdoor detection approach for pre-trained encoders, requiring neither classifier headers nor input labels. We evaluate DECREE on over 400 encoders trojaned under 3 paradigms. We show the effectiveness of our method on image encoders pre-trained on ImageNet and OpenAI's CLIP 400 million image-text pairs. Our method consistently has a high detection accuracy even if we have only limited or no access to the pre-training dataset.
****
Sequential Training of GANs Against GAN-Classifiers Reveals Correlated "Knowledge Gaps" Present Among Independently Trained GAN Instances
Arkanath Pathak, Nicholas Dufour; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24460-24469
Modern Generative Adversarial Networks (GANs) generate realistic images remarkably well. Previous work has demonstrated the feasibility of "GAN-classifiers" that are distinct from the co-trained discriminator, and operate on images generated from a frozen GAN. That such classifiers work at all affirms the existence of "knowledge gaps" (out-of-distribution artifacts across samples) present in GAN training. We iteratively train GAN-classifiers and train GANs that "fool" the classifiers (in an attempt to fill the knowledge gaps), and examine the effect on GAN training dynamics, output quality, and GAN-classifier generalization. We investigate two settings, a small DCGAN architecture trained on low dimensional images (MNIST), and StyleGAN2, a SOTA GAN architecture trained on high dimensional images (FFHQ). We find that the DCGAN is unable to effectively fool a held-out GAN-classifier without compromising the output quality. However, StyleGAN2 can fool held-out classifiers with no change in output quality, and this effect persists over multiple rounds of GAN/classifier training which appears to reveal an ordering over optima in the generator parameter space. Finally, we study different classifier architectures and show that the architecture of the GAN-classifier has a strong influence on the set of its learned artifacts.
****
Lookahead Diffusion Probabilistic Models for Refining Mean Estimation
Guoqiang Zhang, Kenta Niwa, W. Bastiaan Kleijn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1421-1429
We propose lookahead diffusion probabilistic models (LA-DPMs) to exploit the correlation in the outputs of the deep neural networks (DNNs) over subsequent timesteps in diffusion probabilistic models (DPMs) to refine the mean estimation of the conditional Gaussian distributions in the backward process. A typical DPM first obtains an estimate of the original data sample x by feeding the most recent state z_i and index i into the DNN model and then computes the mean vector of the conditional Gaussian distribution for z_ i-1 . We propose to calculate a more accurate estimate for x by performing extrapolation on the two estimates of x that are obtained by feeding (z_ i+1 , i+1) and (z_i, i) into the DNN model. The extrapolation can be easily integrated into the backward process of existing DPMs by introducing an additional connection over two consecutive timesteps, and fine-tuning is not required. Extensive experiments showed that plugging in the additional connection into DDPM, DDIM, DEIS, S-PNDM, and high-order DPM-Solvers leads to a significant performance gain in terms of Frechet inception distance (FID) score. Our implementation is available at https://github.com/guoqiangzhang-x/LA-DPM.
****
TensoIR: Tensorial Inverse Rendering
Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, Hao Su; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 165-174
We propose TensoIR, a novel inverse rendering approach based on tensor factorization and neural fields. Unlike previous works that use purely MLP-based neural fields, thus suffering from low capacity and high computation costs, we extend TensoRF, a state-of-the-art approach for radiance field modeling, to estimate scen

e geometry, surface reflectance, and environment illumination from multi-view images captured under unknown lighting conditions. Our approach jointly achieves radiance field reconstruction and physically-based model estimation, leading to photo-realistic novel view synthesis and relighting. Benefiting from the efficiency and extensibility of the TensoRF-based representation, our method can accurately model secondary shading effects (like shadows and indirect lighting) and generally support input images captured under a single or multiple unknown lighting conditions. The low-rank tensor representation allows us to not only achieve fast and compact reconstruction but also better exploit shared information under an arbitrary number of capturing lighting conditions. We demonstrate the superiority of our method to baseline methods qualitatively and quantitatively on various challenging synthetic and real-world scenes.
****
NIPQ: Noise Proxy-Based Integrated Pseudo-Quantization
Juncheol Shin, Junhyuk So, Sein Park, Seungyeop Kang, Sungjoo Yoo, Eunhyeok Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3852-3861
Straight-through estimator (STE), which enables the gradient flow over the non-differentiable function via approximation, has been favored in studies related to quantization-aware training (QAT). However, STE incurs unstable convergence during QAT, resulting in notable quality degradation in low-precision representation. Recently, pseudo-quantization training has been proposed as an alternative approach to updating the learnable parameters using the pseudo-quantization noise instead of STE. In this study, we propose a novel noise proxy-based integrated pseudo-quantization (NIPQ) that enables unified support of pseudo-quantization for both activation and weight with minimal error by integrating the idea of truncation on the pseudo-quantization framework. NIPQ updates all of the quantization parameters (e.g., bit-width and truncation boundary) as well as the network parameters via gradient descent without STE instability, resulting in greatly-simplified but reliable precision allocation without human intervention. Our extensive experiments show that NIPQ outperforms existing quantization algorithms in various vision and language applications by a large margin.
****
Primitive Generation and Semantic-Related Alignment for Universal Zero-Shot Segmentation
Shuting He, Henghui Ding, Wei Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11238-11247
We study universal zero-shot segmentation in this work to achieve panoptic, instance, and semantic segmentation for novel categories without any training samples. Such zero-shot segmentation ability relies on inter-class relationships in semantic space to transfer the visual knowledge learned from seen categories to unseen ones. Thus, it is desired to well bridge semantic-visual spaces and apply the semantic relationships to visual feature learning. We introduce a generative model to synthesize features for unseen categories, which links semantic and visual spaces as well as address the issue of lack of unseen training data. Furthermore, to mitigate the domain gap between semantic and visual spaces, firstly, we enhance the vanilla generator with learned primitives, each of which contains fine-grained attributes related to categories, and synthesize unseen features by selectively assembling these primitives. Secondly, we propose to disentangle the visual feature into the semantic-related part and the semantic-unrelated part that contains useful visual classification clues but is less relevant to semantic representation. The inter-class relationships of semantic-related visual features are then required to be aligned with those in semantic space, thereby transferring semantic knowledge to visual feature learning. The proposed approach achieves impressively state-of-the-art performance on zero-shot panoptic segmentation, instance segmentation, and semantic segmentation.
****
Long Range Pooling for 3D Large-Scale Scene Understanding
Xiang-Li Li, Meng-Hao Guo, Tai-Jiang Mu, Ralph R. Martin, Shi-Min Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

2023, pp. 10300-10311
Inspired by the success of recent vision transformers and large kernel design in convolutional neural networks (CNNs), in this paper, we analyze and explore essential reasons for their success. We claim two factors that are critical for 3D large-scale scene understanding: a larger receptive field and operations with greater non-linearity. The former is responsible for providing long range contexts and the latter can enhance the capacity of the network. To achieve the above properties, we propose a simple yet effective long range pooling (LRP) module using dilation max pooling, which provides a network with a large adaptive receptive field. LRP has few parameters, and can be readily added to current CNNs. Also, based on LRP, we present an entire network architecture, LRPNet, for 3D understanding. Ablation studies are presented to support our claims, and show that the LRP module achieves better results than large kernel convolution yet with reduced computation, due to its non-linearity. We also demonstrate the superiority of LRPNet on various benchmarks: LRPNet performs the best on ScanNet and surpasses other CNN-based methods on S3DIS and Matterport3D. Code will be avalible at https://github.com/li-xl/LRPNet.
****
Object-Goal Visual Navigation via Effective Exploration of Relations Among Historical Navigation States
Heming Du, Lincheng Li, Zi Huang, Xin Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2563-2573
Object-goal visual navigation aims at steering an agent toward an object via a series of moving steps. Previous works mainly focus on learning informative visual representations for navigation, but overlook the impacts of navigation states on the effectiveness and efficiency of navigation. We observe that high relevance among navigation states will cause navigation inefficiency or failure for existing methods. In this paper, we present a History-inspired Navigation Policy Learning (HiNL) framework to estimate navigation states effectively by exploring relationships among historical navigation states. In HiNL, we propose a History-aware State Estimation (HaSE) module to alleviate the impacts of dominant historical states on the current state estimation. Meanwhile, HaSE also encourages an agent to be alert to the current observation changes, thus enabling the agent to make valid actions. Furthermore, we design a History-based State Regularization (HbSR) to explicitly suppress the correlation among navigation states in training. As a result, our agent can update states more effectively while reducing the correlations among navigation states. Experiments on the artificial platform AI2-THOR (i.e.,, iTHOR and RoboTHOR) demonstrate that HiNL significantly outperforms state-of-the-art methods on both Success Rate and SPL in unseen testing environments.
****
Causally-Aware Intraoperative Imputation for Overall Survival Time Prediction
Xiang Li, Xuelin Qian, Litian Liang, Lingjie Kong, Qiaole Dong, Jiejun Chen, Dingxia Liu, Xiuzhong Yao, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15681-15690
Previous efforts in vision community are mostly made on learning good representations from visual patterns. Beyond this, this paper emphasizes the high-level ability of causal reasoning. We thus present a case study of solving the challenging task of Overall Survival (OS) time in primary liver cancers. Critically, the prediction of OS time at the early stage remains challenging, due to the unobvious image patterns of reflecting the OS. To this end, we propose a causal inference system by leveraging the intraoperative attributes and the correlation among them, as an intermediate supervision to bridge the gap between the images and the final OS. Particularly, we build a causal graph, and train the images to estimate the intraoperative attributes for final OS prediction. We present a novel Causally-aware Intraoperative Imputation Model (CAWIM) that can sequentially predict each attribute using its parent nodes in the estimated causal graph. To determine the causal directions, we propose a splitting-voting mechanism, which votes for the direction for each pair of adjacent nodes among multiple predictions obtained via causal discovery from heterogeneity. The practicability and effective

ness of our method are demonstrated by the promising result on liver cancer data set of 361 patients with long-term observations.
****
Probabilistic Knowledge Distillation of Face Ensembles
Jianqing Xu, Shen Li, Ailin Deng, Miao Xiong, Jiaying Wu, Jiaxiang Wu, Shouhong Ding, Bryan Hooi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3489-3498
Mean ensemble (i.e. averaging predictions from multiple models) is a commonly-used technique in machine learning that improves the performance of each individual model. We formalize it as feature alignment for ensemble in open-set face recognition and generalize it into Bayesian Ensemble Averaging (BEA) through the lens of probabilistic modeling. This generalization brings up two practical benefits that existing methods could not provide: (1) the uncertainty of a face image can be evaluated and further decomposed into aleatoric uncertainty and epistemic uncertainty, the latter of which can be used as a measure for out-of-distribution detection of faceness; (2) a BEA statistic provably reflects the aleatoric uncertainty of a face image, acting as a measure for face image quality to improve recognition performance. To inherit the uncertainty estimation capability from BEA without the loss of inference efficiency, we propose BEA-KD, a student model to distill knowledge from BEA. BEA-KD mimics the overall behavior of ensemble members and consistently outperforms SOTA knowledge distillation methods on various challenging benchmarks.
****
Twin Contrastive Learning With Noisy Labels
Zhizhong Huang, Junping Zhang, Hongming Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11661-11670
Learning from noisy data is a challenging task that significantly degenerates the model performance. In this paper, we present TCL, a novel twin contrastive learning model to learn robust representations and handle noisy labels for classification. Specifically, we construct a Gaussian mixture model (GMM) over the representations by injecting the supervised model predictions into GMM to link label-free latent variables in GMM with label-noisy annotations. Then, TCL detects the examples with wrong labels as the out-of-distribution examples by another two-component GMM, taking into account the data distribution. We further propose a cross-supervision with an entropy regularization loss that bootstraps the true targets from model predictions to handle the noisy labels. As a result, TCL can learn discriminative representations aligned with estimated labels through mixup and contrastive learning. Extensive experimental results on several standard benchmarks and real-world datasets demonstrate the superior performance of TCL. In particular, TCL achieves 7.5% improvements on CIFAR-10 with 90% noisy label---an extremely noisy scenario. The source code is available at https://github.com/Hzzone/TCL.
****
TriVol: Point Cloud Rendering via Triple Volumes
Tao Hu, Xiaogang Xu, Ruihang Chu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20732-20741
Existing learning-based methods for point cloud rendering adopt various 3D representations and feature querying mechanisms to alleviate the sparsity problem of point clouds. However, artifacts still appear in the rendered images, due to the challenges in extracting continuous and discriminative 3D features from point clouds. In this paper, we present a dense while lightweight 3D representation, named TriVol, that can be combined with NeRF to render photo-realistic images from point clouds. Our TriVol consists of triple slim volumes, each of which is encoded from the input point cloud. Our representation has two advantages. First, it fuses the respective fields at different scales and thus extracts local and non-local features for discriminative representation. Second, since the volume size is greatly reduced, our 3D decoder can be efficiently inferred, allowing us to increase the resolution of the 3D space to render more point details. Extensive experiments on different benchmarks with varying kinds of scenes/objects demonstrate our framework's effectiveness compared with current approaches. Moreover, o

ur framework has excellent generalization ability to render a category of scenes or objects without fine-tuning.
****
(ML)^{2}P-Encoder: On Exploration of Channel-Class Correlation for Multi-Label Zero-Shot Learning
Ziming Liu, Song Guo, Xiaocheng Lu, Jingcai Guo, Jiewei Zhang, Yue Zeng, Fushuo Huo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23859-23868
Recent studies usually approach multi-label zero-shot learning (MLZSL) with visual-semantic mapping on spatial-class correlation, which can be computationally costly, and worse still, fails to capture fine-grained class-specific semantics. We observe that different channels may usually have different sensitivities on classes, which can correspond to specific semantics. Such an intrinsic channel-class correlation suggests a potential alternative for the more accurate and class-harmonious feature representations. In this paper, our interest is to fully explore the power of channel-class correlation as the unique base for MLZSL. Specifically, we propose a light yet efficient Multi-Label Multi-Layer Perceptron-based Encoder, dubbed (ML)^2P-Encoder, to extract and preserve channel-wise semantics. We reorganize the generated feature maps into several groups, of which each of them can be trained independently with (ML)^2P-Encoder. On top of that, a global group-wise attention module is further designed to build the multi-label specific class relationships among different classes, which eventually fulfills a novel Channel-Class Correlation MLZSL framework (C^3-MLZSL). Extensive experiments on large-scale MLZSL benchmarks including NUS-WIDE and Open-Images-V4 demonstrate the superiority of our model against other representative state-of-the-art models.
****
MeMaHand: Exploiting Mesh-Mano Interaction for Single Image Two-Hand Reconstruction
Congyi Wang, Feida Zhu, Shilei Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 564-573
Existing methods proposed for hand reconstruction tasks usually parameterize a generic 3D hand model or predict hand mesh positions directly. The parametric representations consisting of hand shapes and rotational poses are more stable, while the non-parametric methods can predict more accurate mesh positions. In this paper, we propose to reconstruct meshes and estimate MANO parameters of two hands from a single RGB image simultaneously to utilize the merits of two kinds of hand representations. To fulfill this target, we propose novel Mesh-Mano interaction blocks (MMIBs), which take mesh vertices positions and MANO parameters as two kinds of query tokens. MMIB consists of one graph residual block to aggregate local information and two transformer encoders to model long-range dependencies. The transformer encoders are equipped with different asymmetric attention masks to model the intra-hand and inter-hand attention, respectively. Moreover, we introduce the mesh alignment refinement module to further enhance the mesh-image alignment. Extensive experiments on the InterHand2.6M benchmark demonstrate promising results over the state-of-the-art hand reconstruction methods.
****
Asymmetric Feature Fusion for Image Retrieval
Hui Wu, Min Wang, Wengang Zhou, Zhenbo Lu, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11082-11092
In asymmetric retrieval systems, models with different capacities are deployed on platforms with different computational and storage resources. Despite the great progress, existing approaches still suffer from a dilemma between retrieval efficiency and asymmetric accuracy due to the low capacity of the lightweight query model. In this work, we propose an Asymmetric Feature Fusion (AFF) paradigm, which advances existing asymmetric retrieval systems by considering the complementarity among different features just at the gallery side. Specifically, it first embeds each gallery image into various features, e.g., local features and global features. Then, a dynamic mixer is introduced to aggregate these features into

a compact embedding for efficient search. On the query side, only a single ligh tweight model is deployed for feature extraction. The query model and dynamic mi xer are jointly trained by sharing a momentum-updated classifier. Notably, the p roposed paradigm boosts the accuracy of asymmetric retrieval without introducing any extra overhead to the query side. Exhaustive experiments on various landmar k retrieval datasets demonstrate the superiority of our paradigm.
****
CREPE: Can Vision-Language Foundation Models Reason Compositionally?
Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, Ranjay Krishna; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2023, pp. 10910-10921
A fundamental characteristic common to both human vision and natural language is their compositional nature. Yet, despite the performance gains contributed by l arge vision and language pretraining, we find that--across 7 architectures train ed with 4 algorithms on massive datasets--they struggle at compositionality. To arrive at this conclusion, we introduce a new compositionality evaluation benchm ark, CREPE, which measures two important aspects of compositionality identified by cognitive science literature: systematicity and productivity. To measure syst ematicity, CREPE consists of a test dataset containing over 370K image-text pair s and three different seen-unseen splits. The three splits are designed to test models trained on three popular training datasets: CC-12M, YFCC-15M, and LAION-4 00M. We also generate 325K, 316K, and 309K hard negative captions for a subset o f the pairs. To test productivity, CREPE contains 17K image-text pairs with nine different complexities plus 278K hard negative captions with atomic, swapping, and negation foils. The datasets are generated by repurposing the Visual Genome scene graphs and region descriptions and applying handcrafted templates and GPT-3. For systematicity, we find that model performance decreases consistently when novel compositions dominate the retrieval set, with Recall@1 dropping by up to 9%. For productivity, models' retrieval success decays as complexity increases, frequently nearing random chance at high complexity. These results hold regardle ss of model and training dataset size.
****
DSFNet: Dual Space Fusion Network for Occlusion-Robust 3D Dense Face Alignment
Heyuan Li, Bo Wang, Yu Cheng, Mohan Kankanhalli, Robby T. Tan; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 4531-4540
Sensitivity to severe occlusion and large view angles limits the usage scenarios of the existing monocular 3D dense face alignment methods. The state-of-the-art 3DMM-based method, directly regresses the model's coefficients, underutilizing the low-level 2D spatial and semantic information, which can actually offer cues for face shape and orientation. In this work, we demonstrate how modeling 3D fa cial geometry in image and model space jointly can solve the occlusion and view angle problems. Instead of predicting the whole face directly, we regress image space features in the visible facial region by dense prediction first. Subsequen tly, we predict our model's coefficients based on the regressed feature of the v isible regions, leveraging the prior knowledge of whole face geometry from the m orphable models to complete the invisible regions. We further propose a fusion n etwork that combines the advantages of both the image and model space prediction s to achieve high robustness and accuracy in unconstrained scenarios. Thanks to the proposed fusion module, our method is robust not only to occlusion and large pitch and roll view angles, which is the benefit of our image space approach, b ut also to noise and large yaw angles, which is the benefit of our model space m ethod. Comprehensive evaluations demonstrate the superior performance of our met hod compared with the state-of-the-art methods. On the 3D dense face alignment t ask, we achieve 3.80% NME on the AFLW2000-3D dataset, which outperforms the stat e-of-the-art method by 5.5%. Code is available at https://github.com/lhyfst/DSFN et.
****
MoStGAN-V: Video Generation With Temporal Motion Styles
Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conferen

ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5652-5661
Video generation remains a challenging task due to spatiotemporal complexity and the requirement of synthesizing diverse motions with temporal consistency. Previous works attempt to generate videos in arbitrary lengths either in an autoregressive manner or regarding time as a continuous signal. However, they struggle to synthesize detailed and diverse motions with temporal coherence and tend to generate repetitive scenes after a few time steps. In this work, we argue that a single time-agnostic latent vector of style-based generator is insufficient to model various and temporally-consistent motions. Hence, we introduce additional time-dependent motion styles to model diverse motion patterns. In addition, a Motion Style Attention modulation mechanism, dubbed as MoStAtt, is proposed to augment frames with vivid dynamics for each specific scale (i.e., layer), which assigns attention score for each motion style w.r.t deconvolution filter weights in the target synthesis layer and softly attends different motion styles for weight modulation. Experimental results show our model achieves state-of-the-art performance on four unconditional 256^2 video synthesis benchmarks trained with only 3 frames per clip and produces better qualitative results with respect to dynamic motions. Code and videos have been made available at https://github.com/xiaoqian-shen/MoStGAN-V.
****
Poly-PC: A Polyhedral Network for Multiple Point Cloud Tasks at Once
Tao Xie, Shiguang Wang, Ke Wang, Linqi Yang, Zhiqiang Jiang, Xingcheng Zhang, Kun Dai, Ruifeng Li, Jian Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1233-1243
In this work, we show that it is feasible to perform multiple tasks concurrently on point cloud with a straightforward yet effective multi-task network. Our framework, Poly-PC, tackles the inherent obstacles (e.g., different model architectures caused by task bias and conflicting gradients caused by multiple dataset domains, etc.) of multi-task learning on point cloud. Specifically, we propose a residual set abstraction (Res-SA) layer for efficient and effective scaling in both width and depth of the network, hence accommodating the needs of various tasks. We develop a weight-entanglement-based one-shot NAS technique to find optimal architectures for all tasks. Moreover, such technique entangles the weights of multiple tasks in each layer to offer task-shared parameters for efficient storage deployment while providing ancillary task-specific parameters for learning task-related features. Finally, to facilitate the training of Poly-PC, we introduce a task-prioritization-based gradient balance algorithm that leverages task prioritization to reconcile conflicting gradients, ensuring high performance for all tasks. Benefiting from the suggested techniques, models optimized by Poly-PC collectively for all tasks keep fewer total FLOPs and parameters and outperform previous methods. We also demonstrate that Poly-PC allows incremental learning and evades catastrophic forgetting when tuned to a new task.
****
HandsOff: Labeled Dataset Generation With No Additional Human Annotations
Austin Xu, Mariya I. Vasileva, Achal Dave, Arjun Seshadri; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7991-8000
Recent work leverages the expressive power of genera- tive adversarial networks (GANs) to generate labeled syn- thetic datasets. These dataset generation methods often require new annotations of synthetic images, which forces practitioners to seek out annotators, curate a set of synthetic images, and ensure the quality of generated labels. We in- troduce the HandsOff framework, a technique capable of producing an unlimited number of synthetic images and cor- responding labels after being trained on less than 50 pre- existing labeled images. Our framework avoids the practi- cal drawbacks of prior work by unifying the field of GAN in- version with dataset generation. We generate datasets with rich pixel-wise labels in multiple challenging domains such as faces, cars, full-body human poses, and urban driving scenes. Our method achieves state-of-the-art performance in semantic segmentation, keypoint detection, and depth es- timation compared to prior dataset generation approaches and transfer learning baselines. We additionally

showcase its ability to address broad challenges in model develop- ment which stem from fixed, hand-annotated datasets, such as the long-tail problem in semantic segmentation. Project page: austinxu87.github.io/handsoff.
****
Semi-Supervised 2D Human Pose Estimation Driven by Position Inconsistency Pseudo Label Correction Module
Linzhi Huang, Yulong Li, Hongbo Tian, Yue Yang, Xiangang Li, Weihong Deng, Jieping Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 693-703
In this paper, we delve into semi-supervised 2D human pose estimation. The previous method ignored two problems: (i) When conducting interactive training between large model and lightweight model, the pseudo label of lightweight model will be used to guide large models. (ii) The negative impact of noise pseudo labels on training. Moreover, the labels used for 2D human pose estimation are relatively complex: keypoint category and keypoint position. To solve the problems mentioned above, we propose a semi-supervised 2D human pose estimation framework driven by a position inconsistency pseudo label correction module (SSPCM). We introduce an additional auxiliary teacher and use the pseudo labels generated by the two teacher model in different periods to calculate the inconsistency score and remove outliers. Then, the two teacher models are updated through interactive training, and the student model is updated using the pseudo labels generated by two teachers. To further improve the performance of the student model, we use the semi-supervised Cut-Occlude based on pseudo keypoint perception to generate more hard and effective samples. In addition, we also proposed a new indoor overhead fisheye human keypoint dataset WEPDTOF-Pose. Extensive experiments demonstrate that our method outperforms the previous best semi-supervised 2D human pose estimation method. We will release the code and dataset at https://github.com/hlz0606/SSPCM.
****
ARKitTrack: A New Diverse Dataset for Tracking Using Mobile RGB-D Data
Haojie Zhao, Junsong Chen, Lijun Wang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5126-5135
Compared with traditional RGB-only visual tracking, few datasets have been constructed for RGB-D tracking. In this paper, we propose ARKitTrack, a new RGB-D tracking dataset for both static and dynamic scenes captured by consumer-grade LiDAR scanners equipped on Apple's iPhone and iPad. ARKitTrack contains 300 RGB-D sequences, 455 targets, and 229.7K video frames in total. Along with the bounding box annotations and frame-level attributes, we also annotate this dataset with 123.9K pixel-level target masks. Besides, the camera intrinsic and camera pose of each frame are provided for future developments. To demonstrate the potential usefulness of this dataset, we further present a unified baseline for both box-level and pixel-level tracking, which integrates RGB features with bird's-eye-view representations to better explore cross-modality 3D geometry. In-depth empirical analysis has verified that the ARKitTrack dataset can significantly facilitate RGB-D tracking and that the proposed baseline method compares favorably against the state of the arts. The source code and dataset will be released.
****
Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks
Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, Furu Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19175-19186
A big convergence of language, vision, and multimodal pretraining is emerging. In this work, we introduce a general-purpose multimodal foundation model BEiT-3, which achieves excellent transfer performance on both vision and vision-language tasks. Specifically, we advance the big convergence from three aspects: backbone architecture, pretraining task, and model scaling up. We use Multiway Transformers for general-purpose modeling, where the modular architecture enables both deep fusion and modality-specific encoding. Based on the shared backbone, we perf

orm masked "language" modeling on images (Imglish), texts (English), and image-text pairs ("parallel sentences") in a unified manner. Experimental results show that BEiT-3 obtains remarkable performance on object detection (COCO), semantic segmentation (ADE20K), image classification (ImageNet), visual reasoning (NLVR2), visual question answering (VQAv2), image captioning (COCO), and cross-modal retrieval (Flickr30K, COCO).
****
Density-Insensitive Unsupervised Domain Adaption on 3D Object Detection
Qianjiang Hu, Daizong Liu, Wei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17556-17566
3D object detection from point clouds is crucial in safety-critical autonomous driving. Although many works have made great efforts and achieved significant progress on this task, most of them suffer from expensive annotation cost and poor transferability to unknown data due to the domain gap. Recently, few works attempt to tackle the domain gap in objects, but still fail to adapt to the gap of varying beam-densities between two domains, which is critical to mitigate the characteristic differences of the LiDAR collectors. To this end, we make the attempt to propose a density-insensitive domain adaption framework to address the density-induced domain gap. In particular, we first introduce Random Beam Re-Sampling (RBRS) to enhance the robustness of 3D detectors trained on the source domain to the varying beam-density. Then, we take this pre-trained detector as the backbone model, and feed the unlabeled target domain data into our newly designed task-specific teacher-student framework for predicting its high-quality pseudo labels. To further adapt the property of density-insensitive into the target domain, we feed the teacher and student branches with the same sample of different densities, and propose an Object Graph Alignment (OGA) module to construct two object-graphs between the two branches for enforcing the consistency in both the attribute and relation of cross-density objects. Experimental results on three widely adopted 3D object detection datasets demonstrate that our proposed domain adaption method outperforms the state-of-the-art methods, especially over varying-density data. Code is available at https://github.com/WoodwindHu/DTS.
****
Efficient Verification of Neural Networks Against LVM-Based Specifications
Harleen Hanspal, Alessio Lomuscio; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3894-3903
The deployment of perception systems based on neural networks in safety critical applications requires assurance on their robustness. Deterministic guarantees on network robustness require formal verification. Standard approaches for verifying robustness analyse invariance to analytically defined transformations, but not the diverse and ubiquitous changes involving object pose, scene viewpoint, occlusions, etc. To this end, we present an efficient approach for verifying specifications definable using Latent Variable Models that capture such diverse changes. The approach involves adding an invertible encoding head to the network to be verified, enabling the verification of latent space sets with minimal reconstruction overhead. We report verification experiments for three classes of proposed latent space specifications, each capturing different types of realistic input variations. Differently from previous work in this area, the proposed approach is relatively independent of input dimensionality and scales to a broad class of deep networks and real-world datasets by mitigating the inefficiency and decoder expressivity dependence in the present state-of-the-art.
****
Learning Action Changes by Measuring Verb-Adverb Textual Relationships
Davide Moltisanti, Frank Keller, Hakan Bilen, Laura Sevilla-Lara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23110-23118
The goal of this work is to understand the way actions are performed in videos. That is, given a video, we aim to predict an adverb indicating a modification applied to the action (e.g. cut "finely"). We cast this problem as a regression task. We measure textual relationships between verbs and adverbs to generate a regression target representing the action change we aim to learn. We test our appro

ach on a range of datasets and achieve state-of-the-art results on both adverb prediction and antonym classification. Furthermore, we outperform previous work when we lift two commonly assumed conditions: the availability of action labels during testing and the pairing of adverbs as antonyms. Existing datasets for adverb recognition are either noisy, which makes learning difficult, or contain actions whose appearance is not influenced by adverbs, which makes evaluation less reliable. To address this, we collect a new high quality dataset: Adverbs in Recipes (AIR). We focus on instructional recipes videos, curating a set of actions that exhibit meaningful visual changes when performed differently. Videos in AIR are more tightly trimmed and were manually reviewed by multiple annotators to ensure high labelling quality. Results show that models learn better from AIR given its cleaner videos. At the same time, adverb prediction on AIR is challenging, demonstrating that there is considerable room for improvement.
****

Feature Aggregated Queries for Transformer-Based Video Object Detectors
Yiming Cui; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6365-6376
Video object detection needs to solve feature degradation situations that rarely happen in the image domain. One solution is to use the temporal information and fuse the features from the neighboring frames. With Transformer-based object detectors getting a better performance on the image domain tasks, recent works began to extend those methods to video object detection. However, those existing Transformer-based video object detectors still follow the same pipeline as those used for classical object detectors, like enhancing the object feature representations by aggregation. In this work, we take a different perspective on video object detection. In detail, we improve the qualities of queries for the Transformer-based models by aggregation. To achieve this goal, we first propose a vanilla query aggregation module that weighted averages the queries according to the features of the neighboring frames. Then, we extend the vanilla module to a more practical version, which generates and aggregates queries according to the features of the input frames. Extensive experimental results validate the effectiveness of our proposed methods: On the challenging ImageNet VID benchmark, when integrated with our proposed modules, the current state-of-the-art Transformer-based object detectors can be improved by more than 2.4% on mAP and 4.2% on AP50.
****

Context-Aware Pretraining for Efficient Blind Image Decomposition
Chao Wang, Zhedong Zheng, Ruijie Quan, Yifan Sun, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18186-18195
In this paper, we study Blind Image Decomposition (BID), which is to uniformly remove multiple types of degradation at once without foreknowing the noise type. There remain two practical challenges: (1) Existing methods typically require massive data supervision, making them infeasible to real-world scenarios. (2) The conventional paradigm usually focuses on mining the abnormal pattern of a superimposed image to separate the noise, which de facto conflicts with the primary image restoration task. Therefore, such a pipeline compromises repairing efficiency and authenticity. In an attempt to solve the two challenges in one go, we propose an efficient and simplified paradigm, called Context-aware Pretraining (CP), with two pretext tasks: mixed image separation and masked image reconstruction. Such a paradigm reduces the annotation demands and explicitly facilitates context-aware feature learning. Assuming the restoration process follows a structure-to-texture manner, we also introduce a Context-aware Pretrained network (CPNet). In particular, CPNet contains two transformer-based parallel encoders, one information fusion module, and one multi-head prediction module. The information fusion module explicitly utilizes the mutual correlation in the spatial-channel dimension, while the multi-head prediction module facilitates texture-guided appearance flow. Moreover, a new sampling loss along with an attribute label constraint is also deployed to make use of the spatial context, leading to high-fidelity image restoration. Extensive experiments on both real and synthetic benchmarks show that our method achieves competitive performance for various BID tasks.

****
Weakly Supervised Posture Mining for Fine-Grained Classification
Zhenchao Tang, Hualin Yang, Calvin Yu-Chian Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23735-23744
Because the subtle differences between the different sub-categories of common visual categories such as bird species, fine-grained classification has been seen as a challenging task for many years. Most previous works focus towards the features in the single discriminative region isolatedly, while neglect the connection between the different discriminative regions in the whole image. However, the relationship between different discriminative regions contains rich posture information and by adding the posture information, model can learn the behavior of the object which attribute to improve the classification performance. In this paper, we propose a novel fine-grained framework named PMRC (posture mining and reverse cross-entropy), which is able to combine with different backbones to good effect. In PMRC, we use the Deep Navigator to generate the discriminative regions from the images, and then use them to construct the graph. We aggregate the graph by message passing and get the classification results. Specifically, in order to force PMRC to learn how to mine the posture information, we design a novel training paradigm, which makes the Deep Navigator and message passing communicate and train together. In addition, we propose the reverse cross-entropy (RCE) and demomenstate that compared to the cross-entropy (CE), RCE can not only promote the accurracy of our model but also generalize to promote the accuracy of other kinds of fine-grained classification models. Experimental results on benchmark datasets confirm that PMRC can achieve state-of-the-art.
****
LAVENDER: Unifying Video-Language Understanding As Masked Language Modeling
Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, Lijuan Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23119-23129
Unified vision-language frameworks have greatly advanced in recent years, most of which adopt an encoder-decoder architecture to unify image-text tasks as sequence-to-sequence generation. However, existing video-language (VidL) models still require task-specific designs in model architecture and training objectives for each task. In this work, we explore a unified VidL framework LAVENDER, where Masked Language Modeling (MLM) is used as the common interface for all pre-training and downstream tasks. Such unification leads to a simplified model architecture, where only a lightweight MLM head, instead of a decoder with much more parameters, is needed on top of the multimodal encoder. Surprisingly, experimental results show that this unified framework achieves competitive performance on 14 VidL benchmarks, covering video question answering, text-to-video retrieval and video captioning. Extensive analyses further demonstrate LAVENDER can (i) seamlessly support all downstream tasks with just a single set of parameter values when multi-task finetuned; (ii) generalize to various downstream tasks with limited training samples; and (iii) enable zero-shot evaluation on video question answering tasks.
****
Decomposed Cross-Modal Distillation for RGB-Based Temporal Action Detection
Pilhyeon Lee, Taeoh Kim, Minho Shim, Dongyoon Wee, Hyeran Byun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2373-2383
Temporal action detection aims to predict the time intervals and the classes of action instances in the video. Despite the promising performance, existing two-stream models exhibit slow inference speed due to their reliance on computationally expensive optical flow. In this paper, we introduce a decomposed cross-modal distillation framework to build a strong RGB-based detector by transferring knowledge of the motion modality. Specifically, instead of direct distillation, we propose to separately learn RGB and motion representations, which are in turn combined to perform action localization. The dual-branch design and the asymmetric training objectives enable effective motion knowledge transfer while preserving

RGB information intact. In addition, we introduce a local attentive fusion to better exploit the multimodal complementarity. It is designed to preserve the local discriminability of the features that is important for action localization. Extensive experiments on the benchmarks verify the effectiveness of the proposed method in enhancing RGB-based action detectors. Notably, our framework is agnostic to backbones and detection heads, bringing consistent gains across different model combinations.
****
PyramidFlow: High-Resolution Defect Contrastive Localization Using Pyramid Normalizing Flow
Jiarui Lei, Xiaobo Hu, Yue Wang, Dong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14143-14152
During industrial processing, unforeseen defects may arise in products due to uncontrollable factors. Although unsupervised methods have been successful in defect localization, the usual use of pre-trained models results in low-resolution outputs, which damages visual performance. To address this issue, we propose PyramidFlow, the first fully normalizing flow method without pre-trained models that enables high-resolution defect localization. Specifically, we propose a latent template-based defect contrastive localization paradigm to reduce intra-class variance, as the pre-trained models do. In addition, PyramidFlow utilizes pyramid-like normalizing flows for multi-scale fusing and volume normalization to help generalization. Our comprehensive studies on MVTecAD demonstrate the proposed method outperforms the comparable algorithms that do not use external priors, even achieving state-of-the-art performance in more challenging BTAD scenarios.
****
On-the-Fly Category Discovery
Ruoyi Du, Dongliang Chang, Kongming Liang, Timothy Hospedales, Yi-Zhe Song, Zhanyu Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11691-11700
Although machines have surpassed humans on visual recognition problems, they are still limited to providing closed-set answers. Unlike machines, humans can cognize novel categories at the first observation. Novel category discovery (NCD) techniques, transferring knowledge from seen categories to distinguish unseen categories, aim to bridge the gap. However, current NCD methods assume a transductive learning and offline inference paradigm, which restricts them to a pre-defined query set and renders them unable to deliver instant feedback. In this paper, we study on-the-fly category discovery (OCD) aimed at making the model instantaneously aware of novel category samples (i.e., enabling inductive learning and streaming inference). We first design a hash coding-based expandable recognition model as a practical baseline. Afterwards, noticing the sensitivity of hash codes to intra-category variance, we further propose a novel Sign-Magnitude dIsentangLEment (SMILE) architecture to alleviate the disturbance it brings. Our experimental results demonstrate the superiority of SMILE against our baseline model and prior art. Our code will be made publicly available. Our code is available at https://github.com/PRIS-CV/On-the-fly-Category-Discovery.
****
A Unified Knowledge Distillation Framework for Deep Directed Graphical Models
Yizhuo Chen, Kaizhao Liang, Zhe Zeng, Shuochao Yao, Huajie Shao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7795-7804
Knowledge distillation (KD) is a technique that transfers the knowledge from a large teacher network to a small student network. It has been widely applied to many different tasks, such as model compression and federated learning. However, existing KD methods fail to generalize to general deep directed graphical models (DGMs) with arbitrary layers of random variables. We refer by deep DGMs to DGMs whose conditional distributions are parameterized by deep neural networks. In this work, we propose a novel unified knowledge distillation framework for deep DGMs on various applications. Specifically, we leverage the reparameterization trick to hide the intermediate latent variables, resulting in a compact DGM. Then we develop a surrogate distillation loss to reduce error accumulation through mu

ltiple layers of random variables. Moreover, we present the connections between our method and some existing knowledge distillation approaches. The proposed framework is evaluated on four applications: data-free hierarchical variational autoencoder (VAE) compression, data-free variational recurrent neural networks (VRNN) compression, data-free Helmholtz Machine (HM) compression, and VAE continual learning. The results show that our distillation method outperforms the baselines in data-free model compression tasks. We further demonstrate that our method significantly improves the performance of KD-based continual learning for data generation. Our source code is available at https://github.com/YizhuoChen99/KD4DGM-CVPR.
****
MAIR: Multi-View Attention Inverse Rendering With 3D Spatially-Varying Lighting Estimation
JunYong Choi, SeokYeong Lee, Haesol Park, Seung-Won Jung, Ig-Jae Kim, Junghyun Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8392-8401
We propose a scene-level inverse rendering framework that uses multi-view images to decompose the scene into geometry, a SVBRDF, and 3D spatially-varying lighting. Because multi-view images provide a variety of information about the scene, multi-view images in object-level inverse rendering have been taken for granted. However, owing to the absence of multi-view HDR synthetic dataset, scene-level inverse rendering has mainly been studied using single-view image. We were able to successfully perform scene-level inverse rendering using multi-view images by expanding OpenRooms dataset and designing efficient pipelines to handle multi-view images, and splitting spatially-varying lighting. Our experiments show that the proposed method not only achieves better performance than single-view-based methods, but also achieves robust performance on unseen real-world scene. Also, our sophisticated 3D spatially-varying lighting volume allows for photorealistic object insertion in any 3D location.
****
DF-Platter: Multi-Face Heterogeneous Deepfake Dataset
Kartik Narayan, Harsh Agarwal, Kartik Thakral, Surbhi Mittal, Mayank Vatsa, Richa Singh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9739-9748
Deepfake detection is gaining significant importance in the research community. While most of the research efforts are focused around high-quality images and videos, deepfake generation algorithms today have the capability to generate low-resolution videos, occluded deepfakes, and multiple-subject deepfakes. In this research, we emulate the real-world scenario of deepfake generation and spreading, and propose the DF-Platter dataset, which contains (i) both low-resolution and high-resolution deepfakes generated using multiple generation techniques and (ii) single-subject and multiple-subject deepfakes, with face images of Indian ethnicity. Faces in the dataset are annotated for various attributes such as gender, age, skin tone, and occlusion. The database is prepared in 116 days with continuous usage of 32 GPUs accounting to 1,800 GB cumulative memory. With over 500 GBs in size, the dataset contains a total of 133,260 videos encompassing three sets. To the best of our knowledge, this is one of the largest datasets containing vast variability and multiple challenges. We also provide benchmark results under multiple evaluation settings using popular and state-of-the-art deepfake detection models. Further, benchmark results under c23 and c40 compression are provided. The results demonstrate a significant performance reduction in the deepfake detection task on low-resolution deepfakes and show that the existing techniques fail drastically on multiple-subject deepfakes. It is our assertion that this database will improve the state-of-the-art by extending the capabilities of deepfake detection algorithms to real-world scenarios. The database is available at: http://iab-rubric.org/df-platter-database.
****
Shifted Diffusion for Text-to-Image Generation
Yufan Zhou, Bingchen Liu, Yizhe Zhu, Xiao Yang, Changyou Chen, Jinhui Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP

R), 2023, pp. 10157-10166
We present Corgi, a novel method for text-to-image generation. Corgi is based on
 our proposed shifted diffusion model, which achieves better image embedding gen
eration from input text. Different from the baseline diffusion model used in DAL
L-E 2, our method seamlessly encodes prior knowledge of the pre-trained CLIP mod
el in its diffusion process by designing a new initialization distribution and a
 new transition step of the diffusion. Compared to the strong DALL-E 2 baseline,
 our method performs better in generating image embedding from the text in terms
 of both efficiency and effectiveness, which consequently results in better text
-to-image generation. Extensive large-scale experiments are conducted and evalua
ted in terms of both quantitative measures and human evaluation, indicating a st
ronger generation ability of our method compared to existing ones. Furthermore,
our model enables semi-supervised and language-free training for text-to-image g
eneration, where only part or none of the images in the training dataset have an
 associated caption. Trained with only 1.7% of the images being captioned, our s
emi-supervised model obtains FID results comparable to DALL-E 2 on zero-shot tex
t-to-image generation evaluated on MS-COCO. Corgi also achieves new state-of-the
-art results across different datasets on downstream language-free text-to-image
 generation tasks, outperforming the previous method, Lafite, by a large margin.
****
Robust Unsupervised StyleGAN Image Restoration
Yohan Poirier-Ginter, Jean-François Lalonde; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22292-22301
GAN-based image restoration inverts the generative process to repair images corr
upted by known degradations. Existing unsupervised methods must carefully be tun
ed for each task and degradation level. In this work, we make StyleGAN image res
toration robust: a single set of hyperparameters works across a wide range of de
gradation levels. This makes it possible to handle combinations of several degra
dations, without the need to retune. Our proposed approach relies on a 3-phase p
rogressive latent space extension and a conservative optimizer, which avoids the
 need for any additional regularization terms. Extensive experiments demonstrate
 robustness on inpainting, upsampling, denoising, and deartifacting at varying d
egradations levels, outperforming other StyleGAN-based inversion techniques. Our
 approach also favorably compares to diffusion-based restoration by yielding muc
h more realistic inversion results. Code will be released upon publication.
****
Blemish-Aware and Progressive Face Retouching With Limited Paired Data
Lianxin Xie, Wen Xue, Zhen Xu, Si Wu, Zhiwen Yu, Hau San Wong; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 5599-5608
Face retouching aims to remove facial blemishes, while at the same time maintain
ing the textual details of a given input image. The main challenge lies in disti
nguishing blemishes from the facial characteristics, such as moles. Training an
image-to-image translation network with pixel-wise supervision suffers from the
problem of expensive paired training data, since professional retouching needs s
pecialized experience and is time-consuming. In this paper, we propose a Blemish
-aware and Progressive Face Retouching model, which is referred to as BPFRe. Our
 framework can be partitioned into two manageable stages to perform progressive
blemish removal. Specifically, an encoder-decoder-based module learns to coarsel
y remove the blemishes at the first stage, and the resulting intermediate featur
es are injected into a generator to enrich local detail at the second stage. We
find that explicitly suppressing the blemishes can contribute to an effective co
llaboration among the components. Toward this end, we incorporate an attention m
odule, which learns to infer a blemish-aware map and further determine the corre
sponding weights, which are then used to refine the intermediate features transf
erred from the encoder to the decoder, and from the decoder to the generator. Th
erefore, BPFRe is able to deliver significant performance gains on a wide range
of face retouching tasks. It is worth noting that we reduce the dependence of BP
FRe on paired training samples by imposing effective regularization on unpaired
ones.

****
Event-Based Frame Interpolation With Ad-Hoc Deblurring
Lei Sun, Christos Sakaridis, Jingyun Liang, Peng Sun, Jiezhang Cao, Kai Zhang, Qi Jiang, Kaiwei Wang, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18043-18052
The performance of video frame interpolation is inherently correlated with the ability to handle motion in the input scene. Even though previous works recognize the utility of asynchronous event information for this task, they ignore the fact that motion may or may not result in blur in the input video to be interpolated, depending on the length of the exposure time of the frames and the speed of the motion, and assume either that the input video is sharp, restricting themselves to frame interpolation, or that it is blurry, including an explicit, separate deblurring stage before interpolation in their pipeline. We instead propose a general method for event-based frame interpolation that performs deblurring ad-hoc and thus works both on sharp and blurry input videos. Our model consists in a bidirectional recurrent network that naturally incorporates the temporal dimension of interpolation and fuses information from the input frames and the events adaptively based on their temporal proximity. In addition, we introduce a novel real-world high-resolution dataset with events and color videos which provides a challenging evaluation setting for the examined task. Extensive experiments on the standard GoPro benchmark and on our dataset show that our network consistently outperforms previous state-of-the-art methods on frame interpolation, single image deblurring and the joint task of interpolation and deblurring. Our code and dataset will be available at https://github.com/AHupuJR/REFID.
****
OvarNet: Towards Open-Vocabulary Object Attribute Recognition
Keyan Chen, Xiaolong Jiang, Yao Hu, Xu Tang, Yan Gao, Jianqi Chen, Weidi Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23518-23527
In this paper, we consider the problem of simultaneously detecting objects and inferring their visual attributes in an image, even for those with no manual annotations provided at the training stage, resembling an open-vocabulary scenario. To achieve this goal, we make the following contributions: (i) we start with a naive two-stage approach for open-vocabulary object detection and attribute classification, termed CLIP-Attr. The candidate objects are first proposed with an offline RPN and later classified for semantic category and attributes; (ii) we combine all available datasets and train with a federated strategy to finetune the CLIP model, aligning the visual representation with attributes, additionally, we investigate the efficacy of leveraging freely available online image-caption pairs under weakly supervised learning; (iii) in pursuit of efficiency, we train a Faster-RCNN type model end-to-end with knowledge distillation, that performs class-agnostic object proposals and classification on semantic categories and attributes with classifiers generated from a text encoder; Finally, (iv) we conduct extensive experiments on VAW, MS-COCO, LSA, and OVAD datasets, and show that recognition of semantic category and attributes is complementary for visual scene understanding, i.e., jointly training object detection and attributes prediction largely outperform existing approaches that treat the two tasks independently, demonstrating strong generalization ability to novel attributes and categories.
****
Detecting and Grounding Multi-Modal Media Manipulation
Rui Shao, Tianxing Wu, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6904-6913
Misinformation has become a pressing issue. Fake media, in both visual and textual forms, is widespread on the web. While various deepfake detection and text fake news detection methods have been proposed, they are only designed for single-modality forgery based on binary classification, let alone analyzing and reasoning subtle forgery traces across different modalities. In this paper, we highlight a new research problem for multi-modal fake media, namely Detecting and Grounding Multi-Modal Media Manipulation (DGM^4). DGM^4 aims to not only detect the authenticity of multi-modal media, but also ground the manipulated content (i.e.,

image bounding boxes and text tokens), which requires deeper reasoning of multi-modal media manipulation. To support a large-scale investigation, we construct the first DGM^4 dataset, where image-text pairs are manipulated by various approaches, with rich annotation of diverse manipulations. Moreover, we propose a novel HierArchical Multi-modal Manipulation rEasoning tRansformer (HAMMER) to fully capture the fine-grained interaction between different modalities. HAMMER performs 1) manipulation-aware contrastive learning between two uni-modal encoders as shallow manipulation reasoning, and 2) modality-aware cross-attention by multi-modal aggregator as deep manipulation reasoning. Dedicated manipulation detection and grounding heads are integrated from shallow to deep levels based on the interacted multi-modal information. Finally, we build an extensive benchmark and set up rigorous evaluation metrics for this new research problem. Comprehensive experiments demonstrate the superiority of our model; several valuable observations are also revealed to facilitate future research in multi-modal media manipulation.
****
Boosting Detection in Crowd Analysis via Underutilized Output Features
Shaokai Wu, Fengyu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15609-15618
Detection-based methods have been viewed unfavorably in crowd analysis due to their poor performance in dense crowds. However, we argue that the potential of these methods has been underestimated, as they offer crucial information for crowd analysis that is often ignored. Specifically, the area size and confidence score of output proposals and bounding boxes provide insight into the scale and density of the crowd. To leverage these underutilized features, we propose Crowd Hat, a plug-and-play module that can be easily integrated with existing detection models. This module uses a mixed 2D-1D compression technique to refine the output features and obtain the spatial and numerical distribution of crowd-specific information. Based on these features, we further propose region-adaptive NMS thresholds and a decouple-then-align paradigm that address the major limitations of detection-based methods. Our extensive evaluations on various crowd analysis tasks, including crowd counting, localization, and detection, demonstrate the effectiveness of utilizing output features and the potential of detection-based methods in crowd analysis. Our code is available at https://github.com/wskingdom/Crowd-Hat.
****
Human Pose As Compositional Tokens
Zigang Geng, Chunyu Wang, Yixuan Wei, Ze Liu, Houqiang Li, Han Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 660-671
Human pose is typically represented by a coordinate vector of body joints or their heatmap embeddings. While easy for data processing, unrealistic pose estimates are admitted due to the lack of dependency modeling between the body joints. In this paper, we present a structured representation, named Pose as Compositional Tokens (PCT), to explore the joint dependency. It represents a pose by M discrete tokens with each characterizing a sub-structure with several interdependent joints. The compositional design enables it to achieve a small reconstruction error at a low cost. Then we cast pose estimation as a classification task. In particular, we learn a classifier to predict the categories of the M tokens from an image. A pre-learned decoder network is used to recover the pose from the tokens without further post-processing. We show that it achieves better or comparable pose estimation results as the existing methods in general scenarios, yet continues to work well when occlusion occurs, which is ubiquitous in practice. The code and models are publicly available at https://github.com/Gengzigang/PCT.
****
K3DN: Disparity-Aware Kernel Estimation for Dual-Pixel Defocus Deblurring
Yan Yang, Liyuan Pan, Liu Liu, Miaomiao Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13263-13272
The dual-pixel (DP) sensor captures a two-view image pair in a single snapshot by splitting each pixel in half. The disparity occurs in defocus blurred regions

between the two views of the DP pair, while the in-focus sharp regions have zero disparity. This motivates us to propose a K3DN framework for DP pair deblurring, and it has three modules: i) a disparity-aware deblur module. It estimates a disparity feature map, which is used to query a trainable kernel set to estimate a blur kernel that best describes the spatially-varying blur. The kernel is constrained to be symmetrical per the DP formulation. A simple Fourier transform is performed for deblurring that follows the blur model; ii) a reblurring regularization module. It reuses the blur kernel, performs a simple convolution for reblurring, and regularizes the estimated kernel and disparity feature unsupervisedly, in the training stage; iii) a sharp region preservation module. It identifies in-focus regions that correspond to areas with zero disparity between DP images, aims to avoid the introduction of noises during the deblurring process, and improves image restoration performance. Experiments on four standard DP datasets show that the proposed K3DN outperforms state-of-the-art methods, with fewer parameters and flops at the same time.
****

3D Line Mapping Revisited

Shaohui Liu, Yifan Yu, Rémi Pautrat, Marc Pollefeys, Viktor Larsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21445-21455

In contrast to sparse keypoints, a handful of line segments can concisely encode the high-level scene layout, as they often delineate the main structural elements. In addition to offering strong geometric cues, they are also omnipresent in urban landscapes and indoor scenes. Despite their apparent advantages, current line-based reconstruction methods are far behind their point-based counterparts. In this paper we aim to close the gap by introducing LIMAP, a library for 3D line mapping that robustly and efficiently creates 3D line maps from multi-view imagery. This is achieved through revisiting the degeneracy problem of line triangulation, carefully crafted scoring and track building, and exploiting structural priors such as line coincidence, parallelism, and orthogonality. Our code integrates seamlessly with existing point-based Structure-from-Motion methods and can leverage their 3D points to further improve the line reconstruction. Furthermore, as a byproduct, the method is able to recover 3D association graphs between lines and points / vanishing points (VPs). In thorough experiments, we show that LIMAP significantly outperforms existing approaches for 3D line mapping. Our robust 3D line maps also open up new research directions. We show two example applications: visual localization and bundle adjustment, where integrating lines alongside points yields the best results. Code is available at https://github.com/cvg/limap.
****

DartBlur: Privacy Preservation With Detection Artifact Suppression

Baowei Jiang, Bing Bai, Haozhe Lin, Yu Wang, Yuchen Guo, Lu Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16479-16488

Nowadays, privacy issue has become a top priority when training AI algorithms. Machine learning algorithms are expected to benefit our daily life, while personal information must also be carefully protected from exposure. Facial information is particularly sensitive in this regard. Multiple datasets containing facial information have been taken offline, and the community is actively seeking solutions to remedy the privacy issues. Existing methods for privacy preservation can be divided into blur-based and face replacement-based methods. Owing to the advantages of review convenience and good accessibility, blur-based based methods have become a dominant choice in practice. However, blur-based methods would inevitably introduce training artifacts harmful to the performance of downstream tasks. In this paper, we propose a novel De-artifact Blurring(DartBlur) privacy-preserving method, which capitalizes on a DNN architecture to generate blurred faces. DartBlur can effectively hide facial privacy information while detection artifacts are simultaneously suppressed. We have designed four training objectives that particularly aim to improve review convenience and maximize detection artifact suppression. We associate the algorithm with an adversarial training strategy

with a second-order optimization pipeline. Experimental results demonstrate that
DartBlur outperforms the existing face-replacement method from both perspective
s of review convenience and accessibility, and also shows an exclusive advantage
in suppressing the training artifact compared to traditional blur-based methods
. Our implementation is available at https://github.com/JaNg2333/DartBlur.
****
Synthesizing Photorealistic Virtual Humans Through Cross-Modal Disentanglement
Siddarth Ravichandran, Ond■ej Texler, Dimitar Dinev, Hyun Jae Kang; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
23, pp. 4585-4594
Over the last few decades, many aspects of human life have been enhanced with vi
rtual domains, from the advent of digital assistants such as Amazon's Alexa and
Apple's Siri to the latest metaverse efforts of the rebranded Meta. These trends
underscore the importance of generating photorealistic visual depictions of hum
ans. This has led to the rapid growth of so-called deepfake and talking-head gen
eration methods in recent years. Despite their impressive results and popularity
, they usually lack certain qualitative aspects such as texture quality, lips sy
nchronization, or resolution, and practical aspects such as the ability to run i
n real-time. To allow for virtual human avatars to be used in practical scenario
s, we propose an end-to-end framework for synthesizing high-quality virtual huma
n faces capable of speaking with accurate lip motion with a special emphasis on
performance. We introduce a novel network utilizing visemes as an intermediate a
udio representation and a novel data augmentation strategy employing a hierarchi
cal image synthesis approach that allows disentanglement of the different modali
ties used to control the global head motion. Our method runs in real-time, and i
s able to deliver superior results compared to the current state-of-the-art.
****
Test Time Adaptation With Regularized Loss for Weakly Supervised Salient Object
Detection
Olga Veksler; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 7360-7369
It is well known that CNNs tend to overfit to the training data. Test-time adapt
ation is an extreme approach to deal with overfitting: given a test image, the a
im is to adapt the trained model to that image. Indeed nothing can be closer to
the test data than the test image itself. The main difficulty of test-time adapt
ation is that the ground truth is not available. Thus test-time adaptation, whil
e intriguing, applies to only a few scenarios where one can design an effective
loss function that does not require ground truth. We propose the first approach
for test-time Salient Object Detection (SOD) in the context of weak supervision.
Our approach is based on a so called regularized loss function, which can be us
ed for training CNN when pixel precise ground truth is unavailable. Regularized
loss tends to have lower values for the more likely object segments, and thus it
can be used to fine-tune an already trained CNN to a given test image, adapting
to images unseen during training. We develop a regularized loss function partic
ularly suitable for test-time adaptation and show that our approach significantl
y outperforms prior work for weakly supervised SOD.
****
Self-Supervised Pre-Training With Masked Shape Prediction for 3D Scene Understan
ding
Li Jiang, Zetong Yang, Shaoshuai Shi, Vladislav Golyanik, Dengxin Dai, Bernt Sch
iele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2023, pp. 1168-1178
Masked signal modeling has greatly advanced self-supervised pre-training for lan
guage and 2D images. However, it is still not fully explored in 3D scene underst
anding. Thus, this paper introduces Masked Shape Prediction (MSP), a new framewo
rk to conduct masked signal modeling in 3D scenes. MSP uses the essential 3D sem
antic cue, i.e., geometric shape, as the prediction target for masked points. Th
e context-enhanced shape target consisting of explicit shape context and implici
t deep shape feature is proposed to facilitate exploiting contextual cues in sha
pe prediction. Meanwhile, the pre-training architecture in MSP is carefully desi

gned to alleviate the masked shape leakage from point coordinates. Experiments on multiple 3D understanding tasks on both indoor and outdoor datasets demonstrate the effectiveness of MSP in learning good feature representations to consistently boost downstream performance.
****
Efficient and Explicit Modelling of Image Hierarchies for Image Restoration
Yawei Li, Yuchen Fan, Xiaoyu Xiang, Denis Demandolx, Rakesh Ranjan, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18278-18289
The aim of this paper is to propose a mechanism to efficiently and explicitly model image hierarchies in the global, regional, and local range for image restoration. To achieve that, we start by analyzing two important properties of natural images including cross-scale similarity and anisotropic image features. Inspired by that, we propose the anchored stripe self-attention which achieves a good balance between the space and time complexity of self-attention and the modelling capacity beyond the regional range. Then we propose a new network architecture dubbed GRL to explicitly model image hierarchies in the Global, Regional, and Local range via anchored stripe self-attention, window self-attention, and channel attention enhanced convolution. Finally, the proposed network is applied to 7 image restoration types, covering both real and synthetic settings. The proposed method sets the new state-of-the-art for several of those. Code will be available at https://github.com/ofsoundof/GRL-Image-Restoration.git.
****
Guiding Pseudo-Labels With Uncertainty Estimation for Source-Free Unsupervised Domain Adaptation
Mattia Litrico, Alessio Del Bue, Pietro Morerio; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7640-7650
Standard Unsupervised Domain Adaptation (UDA) methods assume the availability of both source and target data during the adaptation. In this work, we investigate Source-free Unsupervised Domain Adaptation (SF-UDA), a specific case of UDA where a model is adapted to a target domain without access to source data. We propose a novel approach for the SF-UDA setting based on a loss reweighting strategy that brings robustness against the noise that inevitably affects the pseudo-labels. The classification loss is reweighted based on the reliability of the pseudo-labels that is measured by estimating their uncertainty. Guided by such reweighting strategy, the pseudo-labels are progressively refined by aggregating knowledge from neighbouring samples. Furthermore, a self-supervised contrastive framework is leveraged as a target space regulariser to enhance such knowledge aggregation. A novel negative pairs exclusion strategy is proposed to identify and exclude negative pairs made of samples sharing the same class, even in presence of some noise in the pseudo-labels. Our method outperforms previous methods on three major benchmarks by a large margin. We set the new SF-UDA state-of-the-art on VisDA-C and DomainNet with a performance gain of +1.8% on both benchmarks and on PACS with +12.3% in the single-source setting and +6.6% in multi-target adaptation. Additional analyses demonstrate that the proposed approach is robust to the noise, which results in significantly more accurate pseudo-labels compared to state-of-the-art approaches.
****
HuManiFlow: Ancestor-Conditioned Normalising Flows on SO(3) Manifolds for Human Pose and Shape Distribution Estimation
Akash Sengupta, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4779-4789
Monocular 3D human pose and shape estimation is an ill-posed problem since multiple 3D solutions can explain a 2D image of a subject. Recent approaches predict a probability distribution over plausible 3D pose and shape parameters conditioned on the image. We show that these approaches exhibit a trade-off between three key properties: (i) accuracy - the likelihood of the ground-truth 3D solution under the predicted distribution, (ii) sample-input consistency - the extent to which 3D samples from the predicted distribution match the visible 2D image evidence, and (iii) sample diversity - the range of plausible 3D solutions modelled b

y the predicted distribution. Our method, HuManiFlow, predicts simultaneously ac curate, consistent and diverse distributions. We use the human kinematic tree to factorise full body pose into ancestor-conditioned per-body-part pose distribut ions in an autoregressive manner. Per-body-part distributions are implemented us ing normalising flows that respect the manifold structure of SO(3), the Lie grou p of per-body-part poses. We show that ill-posed, but ubiquitous, 3D point estim ate losses reduce sample diversity, and employ only probabilistic training losse s. HuManiFlow outperforms state-of-the-art probabilistic approaches on the 3DPW and SSP-3D datasets.
****
DKT: Diverse Knowledge Transfer Transformer for Class Incremental Learning
Xinyuan Gao, Yuhang He, Songlin Dong, Jie Cheng, Xing Wei, Yihong Gong; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023, pp. 24236-24245
Deep neural networks suffer from catastrophic forgetting in class incremental le arning, where the classification accuracy of old classes drastically deteriorate s when the networks learn the knowledge of new classes. Many works have been pro posed to solve the class incremental learning problem. However, most of them eit her suffer from serious catastrophic forgetting and stability-plasticity dilemma or need too many extra parameters and computations. To meet the challenge, we p ropose a novel framework, Diverse Knowledge Transfer Transformer (DKT). which co ntains two novel knowledge transfers based on the attention mechanism to transfe r the task-general knowledge and task-specific knowledge to the current task to alleviate catastrophic forgetting. Besides, we propose a duplex classifier to ad dress the stability-plasticity dilemma, and a novel loss function to cluster the same categories in feature space and discriminate the features between old and new tasks to force the task specific knowledge to be more diverse. Our method ne eds only a few extra parameters, which are negligible, to tackle the increasing number of tasks. We conduct comprehensive experimental results on CIFAR100, Imag eNet100/1000 datasets. The experiment results show that our method outperforms o ther competitive methods and achieves state-of-the-art performance.
****
LipFormer: High-Fidelity and Generalizable Talking Face Generation With a Pre-Le arned Facial Codebook
Jiayu Wang, Kang Zhao, Shiwei Zhang, Yingya Zhang, Yujun Shen, Deli Zhao, Jingre n Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2023, pp. 13844-13853
Generating a talking face video from the input audio sequence is a practical yet challenging task. Most existing methods either fail to capture fine facial deta ils or need to train a specific model for each identity. We argue that a codeboo k pre-learned on high-quality face images can serve as a useful prior that facil itates high-fidelity and generalizable talking head synthesis. Thanks to the str ong capability of the codebook in representing face textures, we simplify the ta lking face generation task as finding proper lip-codes to characterize the varia tion of lips during a portrait talking. To this end, we propose LipFormer, a tra nsformer-based framework, to model the audio-visual coherence and predict the li p-codes sequence based on the input audio features. We further introduce an adap tive face warping module, which helps warp the reference face to the target pose in the feature space, to alleviate the difficulty of lip-code prediction under different poses. By this means, LipFormer can make better use of the pre-learned priors in images and is robust to posture change. Extensive experiments show th at LipFormer can produce more realistic talking face videos compared to previous methods and faithfully generalize to unseen identities.
****
Generalizable Local Feature Pre-Training for Deformable Shape Analysis
Souhaib Attaiki, Lei Li, Maks Ovsjanikov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13650-13661
Transfer learning is fundamental for addressing problems in settings with little training data. While several transfer learning approaches have been proposed in 3D, unfortunately, these solutions typically operate on an entire 3D object or

even scene-level and thus, as we show, fail to generalize to new classes, such as deformable organic shapes. In addition, there is currently a lack of understanding of what makes pre-trained features transferable across significantly different 3D shape categories. In this paper, we make a step toward addressing these challenges. First, we analyze the link between feature locality and transferability in tasks involving deformable 3D objects, while also comparing different backbones and losses for local feature pre-training. We observe that with proper training, learned features can be useful in such tasks, but, crucially, only with an appropriate choice of the receptive field size. We then propose a differentiable method for optimizing the receptive field within 3D transfer learning. Jointly, this leads to the first learnable features that can successfully generalize to unseen classes of 3D shapes such as humans and animals. Our extensive experiments show that this approach leads to state-of-the-art results on several downstream tasks such as segmentation, shape correspondence, and classification. Our code is available at https://github.com/pvnieo/vader.
****

TarViS: A Unified Approach for Target-Based Video Segmentation

Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, Bastian Leibe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18738-18748

The general domain of video segmentation is currently fragmented into different tasks spanning multiple benchmarks. Despite rapid progress in the state-of-the-art, current methods are overwhelmingly task-specific and cannot conceptually generalize to other tasks. Inspired by recent approaches with multi-task capability, we propose TarViS: a novel, unified network architecture that can be applied to any task that requires segmenting a set of arbitrarily defined 'targets' in video. Our approach is flexible with respect to how tasks define these targets, since it models the latter as abstract 'queries' which are then used to predict pixel-precise target masks. A single TarViS model can be trained jointly on a collection of datasets spanning different tasks, and can hot-swap between tasks during inference without any task-specific retraining. To demonstrate its effectiveness, we apply TarViS to four different tasks, namely Video Instance Segmentation (VIS), Video Panoptic Segmentation (VPS), Video Object Segmentation (VOS) and Point Exemplar-guided Tracking (PET). Our unified, jointly trained model achieves state-of-the-art performance on 5/7 benchmarks spanning these four tasks, and competitive performance on the remaining two. Code and model weights are available at: https://github.com/Ali2500/TarViS
****

Progressive Random Convolutions for Single Domain Generalization

Seokeon Choi, Debasmit Das, Sungha Choi, Seunghan Yang, Hyunsin Park, Sungrack Yun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10312-10322

Single domain generalization aims to train a generalizable model with only one source domain to perform well on arbitrary unseen target domains. Image augmentation based on Random Convolutions (RandConv), consisting of one convolution layer randomly initialized for each mini-batch, enables the model to learn generalizable visual representations by distorting local textures despite its simple and lightweight structure. However, RandConv has structural limitations in that the generated image easily loses semantics as the kernel size increases, and lacks the inherent diversity of a single convolution operation. To solve the problem, we propose a Progressive Random Convolution (Pro-RandConv) method that recursively stacks random convolution layers with a small kernel size instead of increasing the kernel size. This progressive approach can not only mitigate semantic distortions by reducing the influence of pixels away from the center in the theoretical receptive field, but also create more effective virtual domains by gradually increasing the style diversity. In addition, we develop a basic random convolution layer into a random convolution block including deformable offsets and affine transformation to support texture and contrast diversification, both of which are also randomly initialized. Without complex generators or adversarial learning, we demonstrate that our simple yet effective augmentation strategy outperforms

state-of-the-art methods on single domain generalization benchmarks.
****
IDGI: A Framework To Eliminate Explanation Noise From Integrated Gradients
Ruo Yang, Binghui Wang, Mustafa Bilgic; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23725-23734
Integrated Gradients (IG) as well as its variants are well-known techniques for interpreting the decisions of deep neural networks. While IG-based approaches attain state-of-the-art performance, they often integrate noise into their explanation saliency maps, which reduce their interpretability. To minimize the noise, we examine the source of the noise analytically and propose a new approach to reduce the explanation noise based on our analytical findings. We propose the Important Direction Gradient Integration (IDGI) framework, which can be easily incorporated into any IG-based method that uses the Reimann Integration for integrated gradient computation. Extensive experiments with three IG-based methods show that IDGI improves them drastically on numerous interpretability metrics.
****
OPE-SR: Orthogonal Position Encoding for Designing a Parameter-Free Upsampling Module in Arbitrary-Scale Image Super-Resolution
Gaochao Song, Qian Sun, Luo Zhang, Ran Su, Jianfeng Shi, Ying He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10009-10020
Arbitrary-scale image super-resolution (SR) is often tackled using the implicit neural representation (INR) approach, which relies on a position encoding scheme to improve its representation ability. In this paper, we introduce orthogonal position encoding (OPE), an extension of position encoding, and an OPE-Upscale module to replace the INR-based upsampling module for arbitrary-scale image super-resolution. Our OPE-Upscale module takes 2D coordinates and latent code as inputs, just like INR, but does not require any training parameters. This parameter-free feature allows the OPE-Upscale module to directly perform linear combination operations, resulting in continuous image reconstruction and achieving arbitrary-scale image reconstruction. As a concise SR framework, our method is computationally efficient and consumes less memory than state-of-the-art methods, as confirmed by extensive experiments and evaluations. In addition, our method achieves comparable results with state-of-the-art methods in arbitrary-scale image super-resolution. Lastly, we show that OPE corresponds to a set of orthogonal basis, validating our design principle.
****
Implicit Surface Contrastive Clustering for LiDAR Point Clouds
Zaiwei Zhang, Min Bai, Erran Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21716-21725
Self-supervised pretraining on large unlabeled datasets has shown tremendous success on improving the task performance of many computer vision tasks. However, such techniques have not been widely used for outdoor LiDAR point cloud perception due to its scene complexity and wide range. This prevents impactful application from 2D pretraining frameworks. In this paper, we propose ISCC, a new self-supervised pretraining method, core of which are two pretext tasks newly designed for LiDAR point clouds. The first task focuses on learning semantic information by sorting local groups of points in the scene into a globally consistent set of semantically meaningful clusters using contrastive learning. This is augmented with a second task which reasons about precise surfaces of various parts of the scene through implicit surface reconstruction to learn geometric structures. We demonstrate their effectiveness on transfer learning performance on 3D object detection and semantic segmentation in real world LiDAR scenes. We further design an unsupervised semantic grouping task to showcase the highly semantically meaningful features learned by our approach.
****
EC2: Emergent Communication for Embodied Control
Yao Mu, Shunyu Yao, Mingyu Ding, Ping Luo, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6704-6714

Embodied control requires agents to leverage multi-modal pre-training to quickly learn how to act in new environments, where video demonstrations contain visual and motion details needed for low-level perception and control, and language instructions support generalization with abstract, symbolic structures. While recent approaches apply contrastive learning to force alignment between the two modalities, we hypothesize better modeling their complementary differences can lead to more holistic representations for downstream adaption. To this end, we propose Emergent Communication for Embodied Control (EC^2), a novel scheme to pre-train video-language representations for few-shot embodied control. The key idea is to learn an unsupervised "language" of videos via emergent communication, which bridges the semantics of video details and structures of natural language. We learn embodied representations of video trajectories, emergent language, and natural language using a language model, which is then used to finetune a lightweight policy network for downstream control. Through extensive experiments in Metaworld and Franka Kitchen embodied benchmarks, EC^2 is shown to consistently outperform previous contrastive learning methods for both videos and texts as task inputs. Further ablations confirm the importance of the emergent language, which is beneficial for both video and language learning, and significantly superior to using pre-trained video captions. We also present a quantitative and qualitative analysis of the emergent language and discuss future directions toward better understanding and leveraging emergent communication in embodied tasks.
****
Semantic Ray: Learning a Generalizable Semantic Field With Cross-Reprojection Attention
Fangfu Liu, Chubin Zhang, Yu Zheng, Yueqi Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17386-17396
In this paper, we aim to learn a semantic radiance field from multiple scenes that is accurate, efficient and generalizable. While most existing NeRFs target at the tasks of neural scene rendering, image synthesis and multi-view reconstruction, there are a few attempts such as Semantic-NeRF that explore to learn high-level semantic understanding with the NeRF structure. However, Semantic-NeRF simultaneously learns color and semantic label from a single ray with multiple heads, where the single ray fails to provide rich semantic information. As a result, Semantic NeRF relies on positional encoding and needs to train one specific model for each scene. To address this, we propose Semantic Ray (S-Ray) to fully exploit semantic information along the ray direction from its multi-view reprojections. As directly performing dense attention over multi-view reprojected rays would suffer from heavy computational cost, we design a Cross-Reprojection Attention module with consecutive intra-view radial and cross-view sparse attentions, which decomposes contextual information along reprojected rays and cross multiple views and then collects dense connections by stacking the modules. Experiments show that our S-Ray is able to learn from multiple scenes, and it presents strong generalization ability to adapt to unseen scenes.
****
DynamicDet: A Unified Dynamic Architecture for Object Detection
Zhihao Lin, Yongtao Wang, Jinhe Zhang, Xiaojie Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6282-6291
Dynamic neural network is an emerging research topic in deep learning. With adaptive inference, dynamic models can achieve remarkable accuracy and computational efficiency. However, it is challenging to design a powerful dynamic detector, because of no suitable dynamic architecture and exiting criterion for object detection. To tackle these difficulties, we propose a dynamic framework for object detection, named DynamicDet. Firstly, we carefully design a dynamic architecture based on the nature of the object detection task. Then, we propose an adaptive router to analyze the multi-scale information and to decide the inference route automatically. We also present a novel optimization strategy with an exiting criterion based on the detection losses for our dynamic detectors. Last, we present a variable-speed inference strategy, which helps to realize a wide range of accuracy-speed trade-offs with only one dynamic detector. Extensive experiments cond

ucted on the COCO benchmark demonstrate that the proposed DynamicDet achieves ne w state-of-the-art accuracy-speed trade-offs. For instance, with comparable accu racy, the inference speed of our dynamic detector Dy-YOLOv7-W6 surpasses YOLOv7-E6 by 12%, YOLOv7-D6 by 17%, and YOLOv7-E6E by 39%. The code is available at htt ps://github.com/VDIGPKU/DynamicDet.
****
I2MVFormer: Large Language Model Generated Multi-View Document Supervision for Z ero-Shot Image Classification
Muhammad Ferjad Naeem, Muhammad Gul Zain Ali Khan, Yongqin Xian, Muhammad Zeshan Afzal, Didier Stricker, Luc Van Gool, Federico Tombari; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 151 69-15179
Recent works have shown that unstructured text (documents) from online sources c an serve as useful auxiliary information for zero-shot image classification. How ever, these methods require access to a high-quality source like Wikipedia and a re limited to a single source of information. Large Language Models (LLM) traine d on web-scale text show impressive abilities to repurpose their learned knowled ge for a multitude of tasks. In this work, we provide a novel perspective on usi ng an LLM to provide text supervision for a zero-shot image classification model . The LLM is provided with a few text descriptions from different annotators as examples. The LLM is conditioned on these examples to generate multiple text des criptions for each class (referred to as views). Our proposed model, I2MVFormer, learns multi-view semantic embeddings for zero-shot image classification with t hese class views. We show that each text view of a class provides complementary information allowing a model to learn a highly discriminative class embedding. M oreover, we show that I2MVFormer is better at consuming the multi-view text supe rvision from LLM compared to baseline models. I2MVFormer establishes a new state -of-the-art on three public benchmark datasets for zero-shot image classificatio n with unsupervised semantic embeddings.
****
MixSim: A Hierarchical Framework for Mixed Reality Traffic Simulation
Simon Suo, Kelvin Wong, Justin Xu, James Tu, Alexander Cui, Sergio Casas, Raquel Urtasun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9622-9631
The prevailing way to test a self-driving vehicle (SDV) in simulation involves n on-reactive open-loop replay of real world scenarios. However, in order to safel y deploy SDVs to the real world, we need to evaluate them in closed-loop. Toward s this goal, we propose to leverage the wealth of interesting scenarios captured in the real world and make them reactive and controllable to enable closed-loop SDV evaluation in what-if situations. In particular, we present MixSim, a hiera rchical framework for mixed reality traffic simulation. MixSim explicitly models agent goals as routes along the road network and learns a reactive route-condit ional policy. By inferring each agent's route from the original scenario, MixSim can reactively re-simulate the scenario and enable testing different autonomy s ystems under the same conditions. Furthermore, by varying each agent's route, we can expand the scope of testing to what-if situations with realistic variations in agent behaviors or even safety-critical interactions. Our experiments show t hat MixSim can serve as a realistic, reactive, and controllable digital twin of real world scenarios. For more information, please visit the project website: ht tps://waabi.ai/research/mixsim/
****
ORCa: Glossy Objects As Radiance-Field Cameras
Kushagra Tiwary, Akshat Dave, Nikhil Behari, Tzofi Klinghoffer, Ashok Veeraragha van, Ramesh Raskar; Proceedings of the IEEE/CVF Conference on Computer Vision an d Pattern Recognition (CVPR), 2023, pp. 20773-20782
Reflections on glossy objects contain valuable and hidden information about the surrounding environment. By converting these objects into cameras, we can unlock exciting applications, including imaging beyond the camera's field-of-view and from seemingly impossible vantage points, e.g. from reflections on the human eye . However, this task is challenging because reflections depend jointly on object

geometry, material properties, the 3D environment, and the observer's viewing direction. Our approach converts glossy objects with unknown geometry into radiance-field cameras to image the world from the object's perspective. Our key insight is to convert the object surface into a virtual sensor that captures cast reflections as a 2D projection of the 5D environment radiance field visible to and surrounding the object. We show that recovering the environment radiance fields enables depth and radiance estimation from the object to its surroundings in addition to beyond field-of-view novel-view synthesis, i.e. rendering of novel views that are only directly visible to the glossy object present in the scene, but not the observer. Moreover, using the radiance field we can image around occluders caused by close-by objects in the scene. Our method is trained end-to-end on multi-view images of the object and jointly estimates object geometry, diffuse radiance, and the 5D environment radiance field.
****
SECAD-Net: Self-Supervised CAD Reconstruction by Learning Sketch-Extrude Operations
Pu Li, Jianwei Guo, Xiaopeng Zhang, Dong-Ming Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16816-16826
Reverse engineering CAD models from raw geometry is a classic but strenuous research problem. Previous learning-based methods rely heavily on labels due to the supervised design patterns or reconstruct CAD shapes that are not easily editable. In this work, we introduce SECAD-Net, an end-to-end neural network aimed at reconstructing compact and easy-to-edit CAD models in a self-supervised manner. Drawing inspiration from the modeling language that is most commonly used in modern CAD software, we propose to learn 2D sketches and 3D extrusion parameters from raw shapes, from which a set of extrusion cylinders can be generated by extruding each sketch from a 2D plane into a 3D body. By incorporating the Boolean operation (i.e., union), these cylinders can be combined to closely approximate the target geometry. We advocate the use of implicit fields for sketch representation, which allows for creating CAD variations by interpolating latent codes in the sketch latent space. Extensive experiments on both ABC and Fusion 360 datasets demonstrate the effectiveness of our method, and show superiority over state-of-the-art alternatives including the closely related method for supervised CAD reconstruction. We further apply our approach to CAD editing and single-view CAD reconstruction. The code is released at https://github.com/BunnySoCrazy/SECAD-Net.
****
Context-Aware Alignment and Mutual Masking for 3D-Language Pre-Training
Zhao Jin, Munawar Hayat, Yuwei Yang, Yulan Guo, Yinjie Lei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10984-10994
3D visual language reasoning plays an important role in effective human-computer interaction. The current approaches for 3D visual reasoning are task-specific, and lack pre-training methods to learn generic representations that can transfer across various tasks. Despite the encouraging progress in vision-language pre-training for image-text data, 3D-language pre-training is still an open issue due to limited 3D-language paired data, highly sparse and irregular structure of point clouds and ambiguities in spatial relations of 3D objects with viewpoint changes. In this paper, we present a generic 3D-language pre-training approach, that tackles multiple facets of 3D-language reasoning by learning universal representations. Our learning objective constitutes two main parts. 1) Context aware spatial-semantic alignment to establish fine-grained correspondence between point clouds and texts. It reduces relational ambiguities by aligning 3D spatial relationships with textual semantic context. 2) Mutual 3D-Language Masked modeling to enable cross-modality information exchange. Instead of reconstructing sparse 3D points for which language can hardly provide cues, we propose masked proposal reasoning to learn semantic class and mask-invariant representations. Our proposed 3D-language pre-training method achieves promising results once adapted to various downstream tasks, including 3D visual grounding, 3D dense captioning and 3D

question answering. Our codes are available at https://github.com/leolyj/3D-VLP
****
MDL-NAS: A Joint Multi-Domain Learning Framework for Vision Transformer
Shiguang Wang, Tao Xie, Jian Cheng, Xingcheng Zhang, Haijun Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20094-20104
In this work, we introduce MDL-NAS, a unified framework that integrates multiple vision tasks into a manageable supernet and optimizes these tasks collectively under diverse dataset domains. MDL-NAS is storage-efficient since multiple models with a majority of shared parameters can be deposited into a single one. Technically, MDL-NAS constructs a coarse-to-fine search space, where the coarse search space offers various optimal architectures for different tasks while the fine search space provides fine-grained parameter sharing to tackle the inherent obstacles of multi-domain learning. In the fine search space, we suggest two parameter sharing policies, i.e., sequential sharing policy and mask sharing policy. Compared with previous works, such two sharing policies allow for the partial sharing and non-sharing of parameters at each layer of the network, hence attaining real fine-grained parameter sharing. Finally, we present a joint-subnet search algorithm that finds the optimal architecture and sharing parameters for each task within total resource constraints, challenging the traditional practice that downstream vision tasks are typically equipped with backbone networks designed for image classification. Experimentally, we demonstrate that MDL-NAS families fitted with non-hierarchical or hierarchical transformers deliver competitive performance for all tasks compared with state-of-the-art methods while maintaining efficient storage deployment and computation. We also demonstrate that MDL-NAS allows incremental learning and evades catastrophic forgetting when generalizing to a new task.
****
Dual Alignment Unsupervised Domain Adaptation for Video-Text Retrieval
Xiaoshuai Hao, Wanqian Zhang, Dayan Wu, Fei Zhu, Bo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18962-18972
Video-text retrieval is an emerging stream in both computer vision and natural language processing communities, which aims to find relevant videos given text queries. In this paper, we study the notoriously challenging task, i.e., Unsupervised Domain Adaptation Video-text Retrieval (UDAVR), wherein training and testing data come from different distributions. Previous works merely alleviate the domain shift, which however overlook the pairwise misalignment issue in target domain, i.e., there exist no semantic relationships between target videos and texts. To tackle this, we propose a novel method named Dual Alignment Domain Adaptation (DADA). Specifically, we first introduce the cross-modal semantic embedding to generate discriminative source features in a joint embedding space. Besides, we utilize the video and text domain adaptations to smoothly balance the minimization of the domain shifts. To tackle the pairwise misalignment in target domain, we introduce the Dual Alignment Consistency (DAC) to fully exploit the semantic information of both modalities in target domain. The proposed DAC adaptively aligns the video-text pairs which are more likely to be relevant in target domain, enabling that positive pairs are increasing progressively and the noisy ones will potentially be aligned in the later stages. To that end, our method can generate more truly aligned target pairs and ensure the discriminality of target features.Compared with the state-of-the-art methods, DADA achieves 20.18% and 18.61% relative improvements on R@1 under the setting of TGIF->MSRVTT and TGIF->MSVD respectively, demonstrating the superiority of our method.
****
Common Pets in 3D: Dynamic New-View Synthesis of Real-Life Deformable Categories
Samarth Sinha, Roman Shapovalov, Jeremy Reizenstein, Ignacio Rocco, Natalia Neverova, Andrea Vedaldi, David Novotny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4881-4891
Obtaining photorealistic reconstructions of objects from sparse views is inherently ambiguous and can only be achieved by learning suitable reconstruction prior

s. Earlier works on sparse rigid object reconstruction successfully learned such priors from large datasets such as CO3D. In this paper, we extend this approach to dynamic objects. We use cats and dogs as a representative example and introduce Common Pets in 3D (CoP3D), a collection of crowd-sourced videos showing around 4,200 distinct pets. CoP3D is one of the first large-scale datasets for benchmarking non-rigid 3D reconstruction "in the wild". We also propose Tracker-NeRF, a method for learning 4D reconstruction from our dataset. At test time, given a small number of video frames of an unseen sequence, Tracker-NeRF predicts the trajectories and dynamics of the 3D points and generates new views, interpolating viewpoint and time. Results on CoP3D reveal significantly better non-rigid new-view synthesis performance than existing baselines. The data is available on the project webpage: https://cop3d.github.io/.
****
Generalized Decoding for Pixel, Image, and Language
Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee, Jianfeng Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15116-15127
We present X-Decoder, a generalized decoding model that can predict pixel-level segmentation and language tokens seamlessly. X-Decoder takes as input two types of queries: (i) generic non-semantic queries and (ii) semantic queries induced from text inputs, to decode different pixel-level and token-level outputs in the same semantic space. With such a novel design, X-Decoder is the first work that provides a unified way to support all types of image segmentation and a variety of vision-language (VL) tasks. Further, our design enables seamless interactions across tasks at different granularities and brings mutual benefits by learning a common and rich pixel-level visual-semantic understanding space, without any pseudo-labeling. After pretraining on a mixed set of a limited amount of segmentation data and millions of image-text pairs, X-Decoder exhibits strong transferability to a wide range of downstream tasks in both zero-shot and finetuning settings. Notably, it achieves (1) state-of-the-art results on open-vocabulary segmentation and referring segmentation on eight datasets; (2) better or competitive finetuned performance to other generalist and specialist models on segmentation and VL tasks; and (3) flexibility for efficient finetuning and novel task composition. Code, demo, video and visualization are available at: https://x-decoder-vl.github.io.
****
Towards Unified Scene Text Spotting Based on Sequence Generation
Taeho Kil, Seonghyeon Kim, Sukmin Seo, Yoonsik Kim, Daehee Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15223-15232
Sequence generation models have recently made significant progress in unifying various vision tasks. Although some auto-regressive models have demonstrated promising results in end-to-end text spotting, they use specific detection formats while ignoring various text shapes and are limited in the maximum number of text instances that can be detected. To overcome these limitations, we propose a UNIfied scene Text Spotter, called UNITS. Our model unifies various detection formats, including quadrilaterals and polygons, allowing it to detect text in arbitrary shapes. Additionally, we apply starting-point prompting to enable the model to extract texts from an arbitrary starting point, thereby extracting more texts beyond the number of instances it was trained on. Experimental results demonstrate that our method achieves competitive performance compared to state-of-the-art methods. Further analysis shows that UNITS can extract a larger number of texts than it was trained on. We provide the code for our method at https://github.com/clovaai/units.
****
Normal-Guided Garment UV Prediction for Human Re-Texturing
Yasamin Jafarian, Tuanfeng Y. Wang, Duygu Ceylan, Jimei Yang, Nathan Carr, Yi Zhou, Hyun Soo Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4627-4636

Clothes undergo complex geometric deformations, which lead to appearance changes
. To edit human videos in a physically plausible way, a texture map must take in
to account not only the garment transformation induced by the body movements and
 clothes fitting, but also its 3D fine-grained surface geometry. This poses, how
ever, a new challenge of 3D reconstruction of dynamic clothes from an image or a
 video. In this paper, we show that it is possible to edit dressed human images
and videos without 3D reconstruction. We estimate a geometry aware texture map b
etween the garment region in an image and the texture space, a.k.a, UV map. Our
UV map is designed to preserve isometry with respect to the underlying 3D surfac
e by making use of the 3D surface normals predicted from the image. Our approach
 captures the underlying geometry of the garment in a self-supervised way, requi
ring no ground truth annotation of UV maps and can be readily extended to predic
t temporally coherent UV maps. We demonstrate that our method outperforms the st
ate-of-the-art human UV map estimation approaches on both real and synthetic dat
a.
****
Learning Compact Representations for LiDAR Completion and Generation
Yuwen Xiong, Wei-Chiu Ma, Jingkang Wang, Raquel Urtasun; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 107
4-1083
LiDAR provides accurate geometric measurements of the 3D world. Unfortunately, d
ense LiDARs are very expensive and the point clouds captured by low-beam LiDAR a
re often sparse. To address these issues, we present UltraLiDAR, a data-driven f
ramework for scene-level LiDAR completion, LiDAR generation, and LiDAR manipulat
ion. The crux of UltraLiDAR is a compact, discrete representation that encodes t
he point cloud's geometric structure, is robust to noise, and is easy to manipul
ate. We show that by aligning the representation of a sparse point cloud to that
 of a dense point cloud, we can densify the sparse point clouds as if they were
captured by a real high-density LiDAR, drastically reducing the cost. Furthermor
e, by learning a prior over the discrete codebook, we can generate diverse, real
istic LiDAR point clouds for self-driving. We evaluate the effectiveness of Ultr
aLiDAR on sparse-to-dense LiDAR completion and LiDAR generation. Experiments sho
w that densifying real-world point clouds with our approach can significantly im
prove the performance of downstream perception systems. Compared to prior art on
 LiDAR generation, our approach generates much more realistic point clouds. Acco
rding to A/B test, over 98.5% of the time human participants prefer our results
over those of previous methods. Please refer to project page https://waabi.ai/re
search/ultralidar/ for more information.
****
Computational Flash Photography Through Intrinsics
Sepideh Sarajian Maralan, Chris Careaga, Yagiz Aksoy; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16654-
16662
Flash is an essential tool as it often serves as the sole controllable light sou
rce in everyday photography. However, the use of flash is a binary decision at t
he time a photograph is captured with limited control over its characteristics s
uch as strength or color. In this work, we study the computational control of th
e flash light in photographs taken with or without flash. We present a physicall
y motivated intrinsic formulation for flash photograph formation and develop fla
sh decomposition and generation methods for flash and no-flash photographs, resp
ectively. We demonstrate that our intrinsic formulation outperforms alternatives
 in the literature and allows us to computationally control flash in in-the-wild
 images.
****
Hubs and Hyperspheres: Reducing Hubness and Improving Transductive Few-Shot Lear
ning With Hyperspherical Embeddings
Daniel J. Trosten, Rwiddhi Chakraborty, Sigurd Løkse, Kristoffer Knutsen Wickstr
øm, Robert Jenssen, Michael C. Kampffmeyer; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7527-7536
Distance-based classification is frequently used in transductive few-shot learni

ng (FSL). However, due to the high-dimensionality of image representations, FSL classifiers are prone to suffer from the hubness problem, where a few points (hubs) occur frequently in multiple nearest neighbour lists of other points. Hubness negatively impacts distance-based classification when hubs from one class appear often among the nearest neighbors of points from another class, degrading the classifier's performance. To address the hubness problem in FSL, we first prove that hubness can be eliminated by distributing representations uniformly on the hypersphere. We then propose two new approaches to embed representations on the hypersphere, which we prove optimize a tradeoff between uniformity and local similarity preservation -- reducing hubness while retaining class structure. Our experiments show that the proposed methods reduce hubness, and significantly improves transductive FSL accuracy for a wide range of classifiers.
****

Improving Graph Representation for Point Cloud Segmentation via Attentive Filtering

Nan Zhang, Zhiyi Pan, Thomas H. Li, Wei Gao, Ge Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1244-1254

Recently, self-attention networks achieve impressive performance in point cloud segmentation due to their superiority in modeling long-range dependencies. However, compared to self-attention mechanism, we find graph convolutions show a stronger ability in capturing local geometry information with less computational cost. In this paper, we employ a hybrid architecture design to construct our Graph Convolution Network with Attentive Filtering (AF-GCN), which takes advantage of both graph convolution and self-attention mechanism. We adopt graph convolutions to aggregate local features in the shallow encoder stages, while in the deeper stages, we propose a self-attention-like module named Graph Attentive Filter (GAF) to better model long-range contexts from distant neighbors. Besides, to further improve graph representation for point cloud segmentation, we employ a Spatial Feature Projection (SFP) module for graph convolutions which helps to handle spatial variations of unstructured point clouds. Finally, a graph-shared down-sampling and up-sampling strategy is introduced to make full use of the graph structures in point cloud processing. We conduct extensive experiments on multiple datasets including S3DIS, ScanNetV2, Toronto-3D, and ShapeNetPart. Experimental results show our AF-GCN obtains competitive performance.
****

SpaText: Spatio-Textual Representation for Controllable Image Generation

Omri Avrahami, Thomas Hayes, Oran Gafni, Sonal Gupta, Yaniv Taigman, Devi Parikh, Dani Lischinski, Ohad Fried, Xi Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18370-18380

Recent text-to-image diffusion models are able to generate convincing results of unprecedented quality. However, it is nearly impossible to control the shapes of different regions/objects or their layout in a fine-grained fashion. Previous attempts to provide such controls were hindered by their reliance on a fixed set of labels. To this end, we present SpaText --- a new method for text-to-image generation using open-vocabulary scene control. In addition to a global text prompt that describes the entire scene, the user provides a segmentation map where each region of interest is annotated by a free-form natural language description. Due to lack of large-scale datasets that have a detailed textual description for each region in the image, we choose to leverage the current large-scale text-to-image datasets and base our approach on a novel CLIP-based spatio-textual representation, and show its effectiveness on two state-of-the-art diffusion models: pixel-based and latent-based. In addition, we show how to extend the classifier-free guidance method in diffusion models to the multi-conditional case and present an alternative accelerated inference algorithm. Finally, we offer several automatic evaluation metrics and use them, in addition to FID scores and a user study, to evaluate our method and show that it achieves state-of-the-art results on image generation with free-form textual scene control.
****

The ObjectFolder Benchmark: Multisensory Learning With Neural and Real Objects

Ruohan Gao, Yiming Dou, Hao Li, Tanmay Agarwal, Jeannette Bohg, Yunzhu Li, Li Fe
i-Fei, Jiajun Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2023, pp. 17276-17286
We introduce the ObjectFolder Benchmark, a benchmark suite of 10 tasks for multi
sensory object-centric learning, centered around object recognition, reconstruct
ion, and manipulation with sight, sound, and touch. We also introduce the Object
Folder Real dataset, including the multisensory measurements for 100 real-world
household objects, building upon a newly designed pipeline for collecting the 3D
 meshes, videos, impact sounds, and tactile readings of real-world objects. For
each task in the ObjectFolder Benchmark, we conduct systematic benchmarking on b
oth the 1,000 multisensory neural objects from ObjectFolder, and the real multis
ensory data from ObjectFolder Real. Our results demonstrate the importance of mu
ltisensory perception and reveal the respective roles of vision, audio, and touc
h for different object-centric learning tasks. By publicly releasing our dataset
 and benchmark suite, we hope to catalyze and enable new research in multisensor
y object-centric learning in computer vision, robotics, and beyond. Project page
: https://objectfolder.stanford.edu
****
ScaleFL: Resource-Adaptive Federated Learning With Heterogeneous Clients
Fatih Ilhan, Gong Su, Ling Liu; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2023, pp. 24532-24541
Federated learning (FL) is an attractive distributed learning paradigm supportin
g real-time continuous learning and client privacy by default. In most FL approa
ches, all edge clients are assumed to have sufficient computation capabilities t
o participate in the learning of a deep neural network (DNN) model. However, in
real-life applications, some clients may have severely limited resources and can
 only train a much smaller local model. This paper presents ScaleFL, a novel FL
approach with two distinctive mechanisms to handle resource heterogeneity and pr
ovide an equitable FL framework for all clients. First, ScaleFL adaptively scale
s down the DNN model along width and depth dimensions by leveraging early exits
to find the best-fit models for resource-aware local training on distributed cli
ents. In this way, ScaleFL provides an efficient balance of preserving basic and
 complex features in local model splits with various sizes for joint training wh
ile enabling fast inference for model deployment. Second, ScaleFL utilizes self-
distillation among exit predictions during training to improve aggregation throu
gh knowledge transfer among subnetworks. We conduct extensive experiments on ben
chmark CV (CIFAR-10/100, ImageNet) and NLP datasets (SST-2, AgNews). We demonstr
ate that ScaleFL outperforms existing representative heterogeneous FL approaches
 in terms of global/local model performance and provides inference efficiency, w
ith up to 2x latency and 4x model size reduction with negligible performance dro
p below 2%.
****
X3KD: Knowledge Distillation Across Modalities, Tasks and Stages for Multi-Camer
a 3D Object Detection
Marvin Klingner, Shubhankar Borse, Varun Ravi Kumar, Behnaz Rezaei, Venkatraman
Narayanan, Senthil Yogamani, Fatih Porikli; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13343-13353
Recent advances in 3D object detection (3DOD) have obtained remarkably strong re
sults for LiDAR-based models. In contrast, surround-view 3DOD models based on mu
ltiple camera images underperform due to the necessary view transformation of fe
atures from perspective view (PV) to a 3D world representation which is ambiguou
s due to missing depth information. This paper introduces X3KD, a comprehensive
knowledge distillation framework across different modalities, tasks, and stages
for multi-camera 3DOD. Specifically, we propose cross-task distillation from an
instance segmentation teacher (X-IS) in the PV feature extraction stage providin
g supervision without ambiguous error backpropagation through the view transform
ation. After the transformation, we apply cross-modal feature distillation (X-FD
) and adversarial training (X-AT) to improve the 3D world representation of mult
i-camera features through the information contained in a LiDAR-based 3DOD teache
r. Finally, we also employ this teacher for cross-modal output distillation (X-O

D), providing dense supervision at the prediction stage. We perform extensive ab lations of knowledge distillation at different stages of multi-camera 3DOD. Our final X3KD model outperforms previous state-of-the-art approaches on the nuScene s and Waymo datasets and generalizes to RADAR-based 3DOD. Qualitative results vi deo at https://youtu.be/1do9DPFmr38.
****
PCT-Net: Full Resolution Image Harmonization Using Pixel-Wise Color Transformati ons
Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, Björn Stenger; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5917-5926
In this paper, we present PCT-Net, a simple and general image harmonization meth od that can be easily applied to images at full resolution. The key idea is to l earn a parameter network that uses downsampled input images to predict the param eters for pixel-wise color transforms (PCTs) which are applied to each pixel in the full-resolution image. We show that affine color transforms are both efficie nt and effective, resulting in state-of-the-art harmonization results. Moreover, we explore both CNNs and Transformers as the parameter network and show that Tr ansformers lead to better results. We evaluate the proposed method on the public full-resolution iHarmony4 dataset, which is comprised of four datasets, and sho w a reduction of the foreground MSE (fMSE) and MSE values by more than 20% and a n increase of the PSNR value by 1.4dB while keeping the architecture light-weigh t. In a user study with 20 people, we show that the method achieves a higher B-T score than two other recent methods.
****
Architecture, Dataset and Model-Scale Agnostic Data-Free Meta-Learning
Zixuan Hu, Li Shen, Zhenyi Wang, Tongliang Liu, Chun Yuan, Dacheng Tao; Proceedi ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 2023, pp. 7736-7745
The goal of data-free meta-learning is to learn useful prior knowledge from a co llection of pre-trained models without accessing their training data. However, e xisting works only solve the problem in parameter space, which (i) ignore the fr uitful data knowledge contained in the pre-trained models; (ii) can not scale to large-scale pre-trained models; (iii) can only meta-learn pre-trained models wi th the same network architecture. To address those issues, we propose a unified framework, dubbed PURER, which contains: (1) ePisode cUrriculum inveRsion (ECI) during data-free meta training; and (2) invErsion calibRation following inner lo op (ICFIL) during meta testing. During meta training, we propose ECI to perform pseudo episode training for learning to adapt fast to new unseen tasks. Specific ally, we progressively synthesize a sequence of pseudo episodes by distilling th e training data from each pre-trained model. The ECI adaptively increases the di fficulty level of pseudo episodes according to the real-time feedback of the met a model. We formulate the optimization process of meta training with ECI as an a dversarial form in an end-to-end manner. During meta testing, we further propose a simple plug-and-play supplement--ICFIL--only used during meta testing to narr ow the gap between meta training and meta testing task distribution. Extensive e xperiments in various real-world scenarios show the superior performance of ours .
****
Egocentric Video Task Translation
Zihui Xue, Yale Song, Kristen Grauman, Lorenzo Torresani; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23 10-2320
Different video understanding tasks are typically treated in isolation, and even with distinct types of curated data (e.g., classifying sports in one dataset, t racking animals in another). However, in wearable cameras, the immersive egocent ric perspective of a person engaging with the world around them presents an inte rconnected web of video understanding tasks---hand-object manipulations, navigat ion in the space, or human-human interactions---that unfold continuously, driven by the person's goals. We argue that this calls for a much more unified approac

h. We propose EgoTask Translation (EgoT2), which takes a collection of models optimized on separate tasks and learns to translate their outputs for improved performance on any or all of them at once. Unlike traditional transfer or multi-task learning, EgoT2's "flipped design" entails separate task-specific backbones and a task translator shared across all tasks, which captures synergies between even heterogeneous tasks and mitigates task competition. Demonstrating our model on a wide array of video tasks from Ego4D, we show its advantages over existing transfer paradigms and achieve top-ranked results on four of the Ego4D 2022 benchmark challenges.
****
Rawgment: Noise-Accounted RAW Augmentation Enables Recognition in a Wide Variety of Environments
Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, Takeshi Ohashi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14007-14017
Image recognition models that work in challenging environments (e.g., extremely dark, blurry, or high dynamic range conditions) must be useful. However, creating training datasets for such environments is expensive and hard due to the difficulties of data collection and annotation. It is desirable if we could get a robust model without the need for hard-to-obtain datasets. One simple approach is to apply data augmentation such as color jitter and blur to standard RGB (sRGB) images in simple scenes. Unfortunately, this approach struggles to yield realistic images in terms of pixel intensity and noise distribution due to not considering the non-linearity of Image Signal Processors (ISPs) and noise characteristics of image sensors. Instead, we propose a noise-accounted RAW image augmentation method. In essence, color jitter and blur augmentation are applied to a RAW image before applying non-linear ISP, resulting in realistic intensity. Furthermore, we introduce a noise amount alignment method that calibrates the domain gap in the noise property caused by the augmentation. We show that our proposed noise-accounted RAW augmentation method doubles the image recognition accuracy in challenging environments only with simple training data.
****
Reliable and Interpretable Personalized Federated Learning
Zixuan Qin, Liu Yang, Qilong Wang, Yahong Han, Qinghua Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20422-20431
Federated learning can coordinate multiple users to participate in data training while ensuring data privacy. The collaboration of multiple agents allows for a natural connection between federated learning and collective intelligence. When there are large differences in data distribution among clients, it is crucial for federated learning to design a reliable client selection strategy and an interpretable client communication framework to better utilize group knowledge. Herein, a reliable personalized federated learning approach, termed RIPFL, is proposed and fully interpreted from the perspective of social learning. RIPFL reliably selects and divides the clients involved in training such that each client can use different amounts of social information and more effectively communicate with other clients. Simultaneously, the method effectively integrates personal information with the social information generated by the global model from the perspective of Bayesian decision rules and evidence theory, enabling individuals to grow better with the help of collective wisdom. An interpretable federated learning mind is well scalable, and the experimental results indicate that the proposed method has superior robustness and accuracy than other state-of-the-art federated learning algorithms.
****
Optimal Transport Minimization: Crowd Localization on Density Maps for Semi-Supervised Counting
Wei Lin, Antoni B. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21663-21673
The accuracy of crowd counting in images has improved greatly in recent years due to the development of deep neural networks for predicting crowd density maps.

However, most methods do not further explore the ability to localize people in t
he density map, with those few works adopting simple methods, like finding the l
ocal peaks in the density map. In this paper, we propose the optimal transport m
inimization (OT-M) algorithm for crowd localization with density maps. The objec
tive of OT-M is to find a target point map that has the minimal Sinkhorn distanc
e with the input density map, and we propose an iterative algorithm to compute t
he solution. We then apply OT-M to generate hard pseudo-labels (point maps) for
semi-supervised counting, rather than the soft pseudo-labels (density maps) used
 in previous methods. Our hard pseudo-labels provide stronger supervision, and a
lso enable the use of recent density-to-point loss functions for training. We al
so propose a confidence weighting strategy to give higher weight to the more rel
iable unlabeled data. Extensive experiments show that our methods achieve outsta
nding performance on both crowd localization and semi-supervised counting. Code
is available at https://github.com/Elin24/OT-M.
****
AdamsFormer for Spatial Action Localization in the Future
Hyung-gun Chi, Kwonjoon Lee, Nakul Agarwal, Yi Xu, Karthik Ramani, Chiho Choi; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2023, pp. 17885-17895
Predicting future action locations is vital for applications like human-robot co
llaboration. While some computer vision tasks have made progress in predicting h
uman actions, accurately localizing these actions in future frames remains an ar
ea with room for improvement. We introduce a new task called spatial action loca
lization in the future (SALF), which aims to predict action locations in both ob
served and future frames. SALF is challenging because it requires understanding
the underlying physics of video observations to predict future action locations
accurately. To address SALF, we use the concept of NeuralODE, which models the l
atent dynamics of sequential data by solving ordinary differential equations (OD
E) with neural networks. We propose a novel architecture, AdamsFormer, which ext
ends observed frame features to future time horizons by modeling continuous temp
oral dynamics through ODE solving. Specifically, we employ the Adams method, a m
ulti-step approach that efficiently uses information from previous steps without
 discarding it. Our extensive experiments on UCF101-24 and JHMDB-21 datasets dem
onstrate that our proposed model outperforms existing long-range temporal modeli
ng methods by a significant margin in terms of frame-mAP.
****
Leveraging per Image-Token Consistency for Vision-Language Pre-Training
Yunhao Gou, Tom Ko, Hansi Yang, James Kwok, Yu Zhang, Mingxuan Wang; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 19155-19164
Most existing vision-language pre-training (VLP) approaches adopt cross-modal ma
sked language modeling (CMLM) to learn vision-language associations. However, we
 find that CMLM is insufficient for this purpose according to our observations:
(1) Modality bias: a considerable amount of masked tokens in CMLM can be recover
ed with only the language information, ignoring the visual inputs. (2) Under-uti
lization of the unmasked tokens: CMLM primarily focuses on the masked token but
it cannot simultaneously leverage other tokens to learn vision-language associat
ions. To handle those limitations, we propose EPIC (lEveraging Per Image-Token C
onsistency for vision-language pre-training). In EPIC, for each image-sentence p
air, we mask tokens that are salient to the image (i.e., Saliency-based Masking
Strategy) and replace them with alternatives sampled from a language model (i.e.
, Inconsistent Token Generation Procedure), and then the model is required to de
termine for each token in the sentence whether it is consistent with the image (
i.e., Image-Token Consistency Task). The proposed EPIC method is easily combined
 with pre-training methods. Extensive experiments show that the combination of t
he EPIC method and state-of-the-art pre-training approaches, including ViLT, ALB
EF, METER, and X-VLM, leads to significant improvements on downstream tasks. Our
 coude is released at https://github.com/gyhdog99/epic
****
BITE: Beyond Priors for Improved Three-D Dog Pose Estimation

Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J. Black, Silvia Zuff i; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2023, pp. 8867-8876
We address the problem of inferring the 3D shape and pose of dogs from images. G iven the lack of 3D training data, this problem is challenging, and the best met hods lag behind those designed to estimate human shape and pose. To make progres s, we attack the problem from multiple sides at once. First, we need a good 3D s hape prior, like those available for humans. To that end, we learn a dog-specifi c 3D parametric model, called D-SMAL. Second, existing methods focus on dogs in standing poses because when they sit or lie down, their legs are self occluded a nd their bodies deform. Without access to a good pose prior or 3D data, we need an alternative approach. To that end, we exploit contact with the ground as a fo rm of side information. We consider an existing large dataset of dog images and label any 3D contact of the dog with the ground. We exploit body-ground contact in estimating dog pose and find that it significantly improves results. Third, w e develop a novel neural network architecture to infer and exploit this contact information. Fourth, to make progress, we have to be able to measure it. Current evaluation metrics are based on 2D features like keypoints and silhouettes, whi ch do not directly correlate with 3D errors. To address this, we create a synthe tic dataset containing rendered images of scanned 3D dogs. With these advances, our method recovers significantly better dog shape and pose than the state of th e art, and we evaluate this improvement in 3D. Our code, model and test dataset are publicly available for research purposes at https://bite.is.tue.mpg.de.
****
Equivalent Transformation and Dual Stream Network Construction for Mobile Image Super-Resolution
Jiahao Chao, Zhou Zhou, Hongfan Gao, Jiali Gong, Zhengfeng Yang, Zhenbing Zeng, Lydia Dehbi; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2023, pp. 14102-14111
In recent years, there has been an increasing demand for real-time super-resolut ion networks on mobile devices. To address this issue, many lightweight super-re solution models have been proposed. However, these models still contain time-con suming components that increase inference latency, limiting their real-world app lications on mobile devices. In this paper, we propose a novel model for singlei mage super-resolution based on Equivalent Transformation and Dual Stream network construction (ETDS). ET method is proposed to transform time-consuming operator s into time-friendly ones such as convolution and ReLU on mobile devices. Then, a dual stream network is designed to alleviate redundant parameters yielded from ET and enhance the feature extraction ability. Taking full advantage of the adv ance of ET and the dual stream network structure, we develop the efficient SR mo del ETDS for mobile devices. The experimental results demonstrate that our ETDS achieves superior inference speed and reconstruction quality compared to prior l ightweight SR methods on mobile devices. The code is available at https://github .com/ECNUSR/ETDS.
****
UTM: A Unified Multiple Object Tracking Model With Identity-Aware Feature Enhanc ement
Sisi You, Hantao Yao, Bing-Kun Bao, Changsheng Xu; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21876-218 86
Recently, Multiple Object Tracking has achieved great success, which consists of object detection, feature embedding, and identity association. Existing methods apply the three-step or two-step paradigm to generate robust trajectories, wher e identity association is independent of other components. However, the independ ent identity association results in the identity-aware knowledge contained in th e tracklet not be used to boost the detection and embedding modules. To overcome the limitations of existing methods, we introduce a novel Unified Tracking Mode l (UTM) to bridge those three components for generating a positive feedback loop with mutual benefits. The key insight of UTM is the Identity-Aware Feature Enha ncement (IAFE), which is applied to bridge and benefit these three components by

utilizing the identity-aware knowledge to boost detection and embedding. Formally, IAFE contains the Identity-Aware Boosting Attention (IABA) and the Identity-Aware Erasing Attention (IAEA), where IABA enhances the consistent regions between the current frame feature and identity-aware knowledge, and IAEA suppresses the distracted regions in the current frame feature. With better detections and embeddings, higher-quality tracklets can also be generated. Extensive experiments of public and private detections on three benchmarks demonstrate the robustness of UTM.
****
On the Stability-Plasticity Dilemma of Class-Incremental Learning
Dongwan Kim, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20196-20204
A primary goal of class-incremental learning is to strike a balance between stability and plasticity, where models should be both stable enough to retain knowledge learned from previously seen classes, and plastic enough to learn concepts from new classes. While previous works demonstrate strong performance on class-incremental benchmarks, it is not clear whether their success comes from the models being stable, plastic, or a mixture of both. This paper aims to shed light on how effectively recent class-incremental learning algorithms address the stability-plasticity trade-off. We establish analytical tools that measure the stability and plasticity of feature representations, and employ such tools to investigate models trained with various algorithms on large-scale class-incremental benchmarks. Surprisingly, we find that the majority of class-incremental learning algorithms heavily favor stability over plasticity, to the extent that the feature extractor of a model trained on the initial set of classes is no less effective than that of the final incremental model. Our observations not only inspire two simple algorithms that highlight the importance of feature representation analysis, but also suggest that class-incremental learning approaches, in general, should strive for better feature representation learning.
****
Generalization Matters: Loss Minima Flattening via Parameter Hybridization for Efficient Online Knowledge Distillation
Tianli Zhang, Mengqi Xue, Jiangtao Zhang, Haofei Zhang, Yu Wang, Lechao Cheng, Jie Song, Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20176-20185
Most existing online knowledge distillation(OKD) techniques typically require sophisticated modules to produce diverse knowledge for improving students' generalization ability. In this paper, we strive to fully utilize multi-model settings instead of well-designed modules to achieve a distillation effect with excellent generalization performance. Generally, model generalization can be reflected in the flatness of the loss landscape. Since averaging parameters of multiple models can find flatter minima, we are inspired to extend the process to the sampled convex combinations of multi-student models in OKD. Specifically, by linearly weighting students' parameters in each training batch, we construct a Hybrid-Weight Model(HWM) to represent the parameters surrounding involved students. The supervision loss of HWM can estimate the landscape's curvature of the whole region around students to measure the generalization explicitly. Hence we integrate HWM's loss into students' training and propose a novel OKD framework via parameter hybridization(OKDPH) to promote flatter minima and obtain robust solutions. Considering the redundancy of parameters could lead to the collapse of HWM, we further introduce a fusion operation to keep the high similarity of students. Compared to the state-of-the-art(SOTA) OKD methods and SOTA methods of seeking flat minima, our OKDPH achieves higher performance with fewer parameters, benefiting OKD with lightweight and robust characteristics. Our code is publicly available at https://github.com/tianlizhang/OKDPH.
****
Gaussian Label Distribution Learning for Spherical Image Object Detection
Hang Xu, Xinyuan Liu, Qiang Zhao, Yike Ma, Chenggang Yan, Feng Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1033-1042

Spherical image object detection emerges in many applications from virtual reali
ty to robotics and automatic driving, while many existing detectors use ln-norms
 loss for regression of spherical bounding boxes. There are two intrinsic flaws
for ln-norms loss, i.e., independent optimization of parameters and inconsistenc
y between metric (dominated by IoU) and loss. These problems are common in plana
r image detection but more significant in spherical image detection. Solution fo
r these problems has been extensively discussed in planar image detection by usi
ng IoU loss and related variants. However, these solutions cannot be migrated to
 spherical image object detection due to the undifferentiable of the Spherical I
oU (SphIoU). In this paper, we design a simple but effective regression loss bas
ed on Gaussian Label Distribution Learning (GLDL) for spherical image object det
ection. Besides, we observe that the scale of the object in a spherical image va
ries greatly. The huge differences among objects from different categories make
the sample selection strategy based on SphIoU challenging. Therefore, we propose
 GLDL-ATSS as a better training sample selection strategy for objects of the sph
erical image, which can alleviate the drawback of IoU threshold-based strategy o
f scale-sample imbalance. Extensive results on various two datasets with differe
nt baseline detectors show the effectiveness of our approach.
****
High-Resolution Image Reconstruction With Latent Diffusion Models From Human Bra
in Activity
Yu Takagi, Shinji Nishimoto; Proceedings of the IEEE/CVF Conference on Computer
Vision and Pattern Recognition (CVPR), 2023, pp. 14453-14463
Reconstructing visual experiences from human brain activity offers a unique way
to understand how the brain represents the world, and to interpret the connectio
n between computer vision models and our visual system. While deep generative mo
dels have recently been employed for this task, reconstructing realistic images
with high semantic fidelity is still a challenging problem. Here, we propose a n
ew method based on a diffusion model (DM) to reconstruct images from human brain
 activity obtained via functional magnetic resonance imaging (fMRI). More specif
ically, we rely on a latent diffusion model (LDM) termed Stable Diffusion. This
model reduces the computational cost of DMs, while preserving their high generat
ive performance. We also characterize the inner mechanisms of the LDM by studyin
g how its different components (such as the latent vector Z, conditioning inputs
 C, and different elements of the denoising U-Net) relate to distinct brain func
tions. We show that our proposed method can reconstruct high-resolution images w
ith high fidelity in straightforward fashion, without the need for any additiona
l training and fine-tuning of complex deep-learning models. We also provide a qu
antitative interpretation of different LDM components from a neuroscientific per
spective. Overall, our study proposes a promising method for reconstructing imag
es from human brain activity, and provides a new framework for understanding DMs
. Please check out our webpage at https://sites.google.com/view/stablediffusion-
withbrain/.
****
L-CoIns: Language-Based Colorization With Instance Awareness
Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, Boxin Shi; Proceedings o
f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202
3, pp. 19221-19230
Language-based colorization produces plausible colors consistent with the langua
ge description provided by the user. Recent studies introduce additional annotat
ion to prevent color-object coupling and mismatch issues, but they still have di
fficulty in distinguishing instances corresponding to the same object words. In
this paper, we propose a transformer-based framework to automatically aggregate
similar image patches and achieve instance awareness without any additional know
ledge. By applying our presented luminance augmentation and counter-color loss t
o break down the statistical correlation between luminance and color words, our
model is driven to synthesize colors with better descriptive consistency. We fur
ther collect a dataset to provide distinctive visual characteristics and detaile
d language descriptions for multiple instances in the same image. Extensive expe
riments demonstrate our advantages of synthesizing visually pleasing and descrip

tion-consistent results of instance-aware colorization.
****
On the Effects of Self-Supervision and Contrastive Alignment in Deep Multi-View Clustering
Daniel J. Trosten, Sigurd Løkse, Robert Jenssen, Michael C. Kampffmeyer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23976-23985
Self-supervised learning is a central component in recent approaches to deep multi-view clustering (MVC). However, we find large variations in the development of self-supervision-based methods for deep MVC, potentially slowing the progress of the field. To address this, we present DeepMVC, a unified framework for deep MVC that includes many recent methods as instances. We leverage our framework to make key observations about the effect of self-supervision, and in particular, drawbacks of aligning representations with contrastive learning. Further, we prove that contrastive alignment can negatively influence cluster separability, and that this effect becomes worse when the number of views increases. Motivated by our findings, we develop several new DeepMVC instances with new forms of self-supervision. We conduct extensive experiments and find that (i) in line with our theoretical findings, contrastive alignments decreases performance on datasets with many views; (ii) all methods benefit from some form of self-supervision; and (iii) our new instances outperform previous methods on several datasets. Based on our results, we suggest several promising directions for future research. To enhance the openness of the field, we provide an open-source implementation of DeepMVC, including recent models and our new instances. Our implementation includes a consistent evaluation protocol, facilitating fair and accurate evaluation of methods and components.
****
Activating More Pixels in Image Super-Resolution Transformer
Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22367-22377
Transformer-based methods have shown impressive performance in low-level vision tasks, such as image super-resolution. However, we find that these networks can only utilize a limited spatial range of input information through attribution analysis. This implies that the potential of Transformer is still not fully exploited in existing networks. In order to activate more input pixels for better reconstruction, we propose a novel Hybrid Attention Transformer (HAT). It combines both channel attention and window-based self-attention schemes, thus making use of their complementary advantages of being able to utilize global statistics and strong local fitting capability. Moreover, to better aggregate the cross-window information, we introduce an overlapping cross-attention module to enhance the interaction between neighboring window features. In the training stage, we additionally adopt a same-task pre-training strategy to exploit the potential of the model for further improvement. Extensive experiments show the effectiveness of the proposed modules, and we further scale up the model to demonstrate that the performance of this task can be greatly improved. Our overall method significantly outperforms the state-of-the-art methods by more than 1dB. Codes and models are available at https://github.com/XPixelGroup/HAT.
****
BEV-SAN: Accurate BEV 3D Object Detection via Slice Attention Networks
Xiaowei Chi, Jiaming Liu, Ming Lu, Rongyu Zhang, Zhaoqing Wang, Yandong Guo, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17461-17470
Bird's-Eye-View (BEV) 3D Object Detection is a crucial multi-view technique for autonomous driving systems. Recently, plenty of works are proposed, following a similar paradigm consisting of three essential components, i.e., camera feature extraction, BEV feature construction, and task heads. Among the three components, BEV feature construction is BEV-specific compared with 2D tasks. Existing methods aggregate the multi-view camera features to the flattened grid in order to construct the BEV feature. However, flattening the BEV space along the height dim

ension fails to emphasize the informative features of different heights. For example, the barrier is located at a low height while the truck is located at a high height. In this paper, we propose a novel method named BEV Slice Attention Network (BEV-SAN) for exploiting the intrinsic characteristics of different heights. Instead of flattening the BEV space, we first sample along the height dimension to build the global and local BEV slices. Then, the features of BEV slices are aggregated from the camera features and merged by the attention mechanism. Finally, we fuse the merged local and global BEV features by a transformer to generate the final feature map for task heads. The purpose of local BEV slices is to emphasize informative heights. In order to find them, we further propose a LiDAR-guided sampling strategy to leverage the statistical distribution of LiDAR to determine the heights of local slices. Compared with uniform sampling, LiDAR-guided sampling can determine more informative heights. We conduct detailed experiments to demonstrate the effectiveness of BEV-SAN. Code will be released.
****
The Dark Side of Dynamic Routing Neural Networks: Towards Efficiency Backdoor Injection
Simin Chen, Hanlin Chen, Mirazul Haque, Cong Liu, Wei Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24585-24594
Recent advancements in deploying deep neural networks (DNNs) on resource-constrained devices have generated interest in input-adaptive dynamic neural networks (DyNNs). DyNNs offer more efficient inferences and enable the deployment of DNNs on devices with limited resources, such as mobile devices. However, we have discovered a new vulnerability in DyNNs that could potentially compromise their efficiency. Specifically, we investigate whether adversaries can manipulate DyNNs' computational costs to create a false sense of efficiency. To address this question, we propose EfficFrog, an adversarial attack that injects universal efficiency backdoors in DyNNs. To inject a backdoor trigger into DyNNs, EfficFrog poisons only a minimal percentage of the DyNNs' training data. During the inference phase, EfficFrog can slow down the backdoored DyNNs and abuse the computational resources of systems running DyNNs by adding the trigger to any input. To evaluate EfficFrog, we tested it on three DNN backbone architectures (based on VGG16, MobileNet, and ResNet56) using two popular datasets (CIFAR-10 and Tiny ImageNet). Our results demonstrate that EfficFrog reduces the efficiency of DyNNs on triggered input samples while keeping the efficiency of clean samples almost the same.
****
Better "CMOS" Produces Clearer Images: Learning Space-Variant Blur Estimation for Blind Image Super-Resolution
Xuhai Chen, Jiangning Zhang, Chao Xu, Yabiao Wang, Chengjie Wang, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1651-1661
Most of the existing blind image Super-Resolution (SR) methods assume that the blur kernels are space-invariant. However, the blur involved in real applications are usually space-variant due to object motion, out-of-focus, etc., resulting in severe performance drop of the advanced SR methods. To address this problem, we firstly introduce two new datasets with out-of-focus blur, i.e., NYUv2-BSR and Cityscapes-BSR, to support further researches of blind SR with space-variant blur. Based on the datasets, we design a novel Cross-MOdal fuSion network (CMOS) that estimate both blur and semantics simultaneously, which leads to improved SR results. It involves a feature Grouping Interactive Attention (GIA) module to make the two modals interact more effectively and avoid inconsistency. GIA can also be used for the interaction of other features because of the universality of its structure. Qualitative and quantitative experiments compared with state-of-the-art methods on above datasets and real-world images demonstrate the superiority of our method, e.g., obtaining PSNR/SSIM by +1.91/+0.0048 on NYUv2-BSR than MANet.
****
MixTeacher: Mining Promising Labels With Mixed Scale Teacher for Semi-Supervised Object Detection

Liang Liu, Boshen Zhang, Jiangning Zhang, Wuhao Zhang, Zhenye Gan, Guanzhong Tia
n, Wenbing Zhu, Yabiao Wang, Chengjie Wang; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7370-7379
Scale variation across object instances is one of the key challenges in object d
etection. Although modern detection models have achieved remarkable progress in
dealing with the scale variation, it still brings trouble in the semi-supervised
case. Most existing semi-supervised object detection methods rely on strict con
ditions to filter out high-quality pseudo labels from the network predictions. H
owever, we observe that objects with extreme scale tend to have low confidence,
which makes the positive supervision missing for these objects. In this paper, w
e delve into the scale variation problem, and propose a novel framework by intro
ducing a mixed scale teacher to improve the pseudo labels generation and scale i
nvariant learning. In addition, benefiting from the better predictions from mixe
d scale features, we propose to mine pseudo labels with the score promotion of p
redictions across scales. Extensive experiments on MS COCO and PASCAL VOC benchm
arks under various semi-supervised settings demonstrate that our method achieves
new state-of-the-art performance. The code and models will be made publicly ava
ilable.
****
DARE-GRAM: Unsupervised Domain Adaptation Regression by Aligning Inverse Gram Ma
trices
Ismail Nejjar, Qin Wang, Olga Fink; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2023, pp. 11744-11754
Unsupervised Domain Adaptation Regression (DAR) aims to bridge the domain gap be
tween a labeled source dataset and an unlabelled target dataset for regression p
roblems. Recent works mostly focus on learning a deep feature encoder by minimiz
ing the discrepancy between source and target features. In this work, we present
a different perspective for the DAR problem by analyzing the closed-form ordina
ry least square (OLS) solution to the linear regressor in the deep domain adapta
tion context. Rather than aligning the original feature embedding space, we prop
ose to align the inverse Gram matrix of the features, which is motivated by its
presence in the OLS solution and the Gram matrix's ability to capture the featur
e correlations. Specifically, we propose a simple yet effective DAR method which
leverages the pseudo-inverse low-rank property to align the scale and angle in
a selected subspace generated by the pseudo-inverse Gram matrix of the two domai
ns. We evaluate our method on three domain adaptation regression benchmarks. Exp
erimental results demonstrate that our method achieves state-of-the-art performa
nce. Our code is available at https://github.com/ismailnejjar/DARE-GRAM.
****
Bidirectional Copy-Paste for Semi-Supervised Medical Image Segmentation
Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, Yan Wang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1151
4-11524
In semi-supervised medical image segmentation, there exist empirical mismatch pr
oblems between labeled and unlabeled data distribution. The knowledge learned fr
om the labeled data may be largely discarded if treating labeled and unlabeled d
ata separately or training labeled and unlabeled data in an inconsistent manner.
We propose a straightforward method for alleviating the problem -- copy-pasting
labeled and unlabeled data bidirectionally, in a simple Mean Teacher architectu
re. The method encourages unlabeled data to learn comprehensive common semantics
from the labeled data in both inward and outward directions. More importantly,
the consistent learning procedure for labeled and unlabeled data can largely red
uce the empirical distribution gap. In detail, we copy-paste a random crop from
a labeled image (foreground) onto an unlabeled image (background) and an unlabel
ed image (foreground) onto a labeled image (background), respectively. The two m
ixed images are fed into a Student network. It is trained by the generated super
visory signal via bidirectional copy-pasting between the predictions of the unla
beled images from the Teacher and the label maps of the labeled images. We explo
re several design choices of how to copy-paste to make it more effective for min
imizing empirical distribution gaps between labeled and unlabeled data. We revea

l that the simple mechanism of copy-pasting bidirectionally between labeled and unlabeled data is good enough and the experiments show solid gains (e.g., over 21% Dice improvement on ACDC dataset with 5% labeled data) compared with other state-of-the-arts on various semi-supervised medical image segmentation datasets.
****

Learning Discriminative Representations for Skeleton Based Action Recognition
Huanyu Zhou, Qingjie Liu, Yunhong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10608-10617
Human action recognition aims at classifying the category of human action from a segment of a video. Recently, people have dived into designing GCN-based models to extract features from skeletons for performing this task, because skeleton representations are much more efficient and robust than other modalities such as RGB frames. However, when employing the skeleton data, some important clues like related items are also discarded. It results in some ambiguous actions that are hard to be distinguished and tend to be misclassified. To alleviate this problem, we propose an auxiliary feature refinement head (FR Head), which consists of spatial-temporal decoupling and contrastive feature refinement, to obtain discriminative representations of skeletons. Ambiguous samples are dynamically discovered and calibrated in the feature space. Furthermore, FR Head could be imposed on different stages of GCNs to build a multi-level refinement for stronger supervision. Extensive experiments are conducted on NTU RGB+D, NTU RGB+D 120, and NW-UCLA datasets. Our proposed models obtain competitive results from state-of-the-art methods and can help to discriminate those ambiguous samples. Codes are available at https://github.com/zhysora/FR-Head.
****

NeRF in the Palm of Your Hand: Corrective Augmentation for Robotics via Novel-View Synthesis
Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, Chelsea Finn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17907-17917
Expert demonstrations are a rich source of supervision for training visual robotic manipulation policies, but imitation learning methods often require either a large number of demonstrations or expensive online expert supervision to learn reactive closed-loop behaviors. In this work, we introduce SPARTN (Synthetic Perturbations for Augmenting Robot Trajectories via NeRF): a fully-offline data augmentation scheme for improving robot policies that use eye-in-hand cameras. Our approach leverages neural radiance fields (NeRFs) to synthetically inject corrective noise into visual demonstrations: using NeRFs to generate perturbed viewpoints while simultaneously calculating the corrective actions. This requires no additional expert supervision or environment interaction, and distills the geometric information in NeRFs into a real-time reactive RGB-only policy. In a simulated 6-DoF visual grasping benchmark, SPARTN improves offline success rates by 2.8x over imitation learning without the corrective augmentations and even outperforms some methods that use online supervision. It additionally closes the gap between RGB-only and RGB-D success rates, eliminating the previous need for depth sensors. In real-world 6-DoF robotic grasping experiments from limited human demonstrations, our method improves absolute success rates by 22.5% on average, including objects that are traditionally challenging for depth-based methods.
****

NeuMap: Neural Coordinate Mapping by Auto-Transdecoder for Camera Localization
Shitao Tang, Sicong Tang, Andrea Tagliasacchi, Ping Tan, Yasutaka Furukawa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 929-939
This paper presents an end-to-end neural mapping method for camera localization, dubbed NeuMap, encoding a whole scene into a grid of latent codes, with which a Transformer-based auto-decoder regresses 3D coordinates of query pixels. State-of-the-art feature matching methods require each scene to be stored as a 3D point cloud with per-point features, consuming several gigabytes of storage per scene. While compression is possible, performance drops significantly at high compression rates. Conversely, coordinate regression methods achieve high compression

by storing scene information in a neural network but suffer from reduced robustn
ess. NeuMap combines the advantages of both approaches by utilizing 1) learnable
 latent codes for efficient scene representation and 2) a scene-agnostic Transfo
rmer-based auto-decoder to infer coordinates for query pixels. This scene-agnost
ic network design learns robust matching priors from large-scale data and enable
s rapid optimization of codes for new scenes while keeping the network weights f
ixed. Extensive evaluations on five benchmarks show that NeuMap significantly ou
tperforms other coordinate regression methods and achieves comparable performanc
e to feature matching methods while requiring a much smaller scene representatio
n size. For example, NeuMap achieves 39.1% accuracy in the Aachen night benchmar
k with only 6MB of data, whereas alternative methods require 100MB or several gi
gabytes and fail completely under high compression settings. The codes are avail
able at https://github.com/Tangshitao/NeuMap.
****
AShapeFormer: Semantics-Guided Object-Level Active Shape Encoding for 3D Object
Detection via Transformers
Zechuan Li, Hongshan Yu, Zhengeng Yang, Tongjia Chen, Naveed Akhtar; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 1012-1021
3D object detection techniques commonly follow a pipeline that aggregates predic
ted object central point features to compute candidate points. However, these ca
ndidate points contain only positional information, largely ignoring the object-
level shape information. This eventually leads to sub-optimal 3D object detectio
n. In this work, we propose AShapeFormer, a semantics-guided object-level shape
encoding module for 3D object detection. This is a plug-n-play module that lever
ages multi-head attention to encode object shape information. We also propose sh
ape tokens and object-scene positional encoding to ensure that the shape informa
tion is fully exploited. Moreover, we introduce a semantic guidance sub-module t
o sample more foreground points and suppress the influence of background points
for a better object shape perception. We demonstrate a straightforward enhanceme
nt of multiple existing methods with our AShapeFormer. Through extensive experim
ents on the popular SUN RGB-D and ScanNetV2 dataset, we show that our enhanced m
odels are able to outperform the baselines by a considerable absolute margin of
up to 8.1%. Code will be available at https://github.com/ZechuanLi/AShapeFormer
****
SeSDF: Self-Evolved Signed Distance Field for Implicit 3D Clothed Human Reconstr
uction
Yukang Cao, Kai Han, Kwan-Yee K. Wong; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4647-4657
We address the problem of clothed human reconstruction from a single image or un
calibrated multi-view images. Existing methods struggle with reconstructing deta
iled geometry of a clothed human and often require a calibrated setting for mult
i-view reconstruction. We propose a flexible framework which, by leveraging the
parametric SMPL-X model, can take an arbitrary number of input images to reconst
ruct a clothed human model under an uncalibrated setting. At the core of our fra
mework is our novel self-evolved signed distance field (SeSDF) module which allo
ws the framework to learn to deform the signed distance field (SDF) derived from
 the fitted SMPL-X model, such that detailed geometry reflecting the actual clot
hed human can be encoded for better reconstruction. Besides, we propose a simple
 method for self-calibration of multi-view images via the fitted SMPL-X paramete
rs. This lifts the requirement of tedious manual calibration and largely increas
es the flexibility of our method. Further, we introduce an effective occlusion-a
ware feature fusion strategy to account for the most useful features to reconstr
uct the human model. We thoroughly evaluate our framework on public benchmarks,
demonstrating significant superiority over the state-of-the-arts both qualitativ
ely and quantitatively.
****
Deep Depth Estimation From Thermal Image
Ukcheol Shin, Jinsun Park, In So Kweon; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1043-1053

Robust and accurate geometric understanding against adverse weather conditions is one top prioritized conditions to achieve a high-level autonomy of self-driving cars. However, autonomous driving algorithms relying on the visible spectrum band are easily impacted by weather and lighting conditions. A long-wave infrared camera, also known as a thermal imaging camera, is a potential rescue to achieve high-level robustness. However, the missing necessities are the well-established large-scale dataset and public benchmark results. To this end, in this paper, we first built a large-scale Multi-Spectral Stereo (MS^2) dataset, including stereo RGB, stereo NIR, stereo thermal, and stereo LiDAR data along with GNSS/IMU information. The collected dataset provides about 195K synchronized data pairs taken from city, residential, road, campus, and suburban areas in the morning, daytime, and nighttime under clear-sky, cloudy, and rainy conditions. Secondly, we conduct an exhaustive validation process of monocular and stereo depth estimation algorithms designed on visible spectrum bands to benchmark their performance in the thermal image domain. Lastly, we propose a unified depth network that effectively bridges monocular depth and stereo depth tasks from a conditional random field approach perspective. Our dataset and source code are available at https://github.com/UkcheolShin/MS2-MultiSpectralStereoDataset.
****
Cross-GAN Auditing: Unsupervised Identification of Attribute Level Similarities and Differences Between Pretrained Generative Models
Matthew L. Olson, Shusen Liu, Rushil Anirudh, Jayaraman J. Thiagarajan, Peer-Timo Bremer, Weng-Keen Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7981-7990
Generative Adversarial Networks (GANs) are notoriously difficult to train especially for complex distributions and with limited data. This has driven the need for interpretable tools to audit trained networks, for example, to identify biases or ensure fairness. Existing GAN audit tools are restricted to coarse-grained, model-data comparisons based on summary statistics such as FID or recall. In this paper, we propose an alternative approach that compares a newly developed GAN against a prior baseline. To this end, we introduce Cross-GAN Auditing (xGA) that, given an established "reference" GAN and a newly proposed "client" GAN, jointly identifies semantic attributes that are either common across both GANs, novel to the client GAN, or missing from the client GAN. This provides both users and model developers an intuitive assessment of similarity and differences between GANs. We introduce novel metrics to evaluate attribute-based GAN auditing approaches and use these metrics to demonstrate quantitatively that xGA outperforms baseline approaches. We also include qualitative results that illustrate the common, novel and missing attributes identified by xGA from GANs trained on a variety of image datasets.
****
Building Rearticulable Models for Arbitrary 3D Objects From 4D Point Clouds
Shaowei Liu, Saurabh Gupta, Shenlong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21138-21147
We build rearticulable models for arbitrary everyday man-made objects containing an arbitrary number of parts that are connected together in arbitrary ways via 1-degree-of-freedom joints. Given point cloud videos of such everyday objects, our method identifies the distinct object parts, what parts are connected to what other parts, and the properties of the joints connecting each part pair. We do this by jointly optimizing the part segmentation, transformation, and kinematics using a novel energy minimization framework. Our inferred animatable models, enables retargeting to novel poses with sparse point correspondences guidance. We test our method on a new articulating robot dataset and the Sapiens dataset with common daily objects. Experiments show that our method outperforms two leading prior works on various metrics.
****
Backdoor Defense via Adaptively Splitting Poisoned Dataset
Kuofeng Gao, Yang Bai, Jindong Gu, Yong Yang, Shu-Tao Xia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4005-4014

Backdoor defenses have been studied to alleviate the threat of deep neural networks (DNNs) being backdoor attacked and thus maliciously altered. Since DNNs usually adopt some external training data from an untrusted third party, a robust backdoor defense strategy during the training stage is of importance. We argue that the core of training-time defense is to select poisoned samples and to handle them properly. In this work, we summarize the training-time defenses from a unified framework as splitting the poisoned dataset into two data pools. Under our framework, we propose an adaptively splitting dataset-based defense (ASD). Concretely, we apply loss-guided split and meta-learning-inspired split to dynamically update two data pools. With the split clean data pool and polluted data pool, ASD successfully defends against backdoor attacks during training. Extensive experiments on multiple benchmark datasets and DNN models against six state-of-the-art backdoor attacks demonstrate the superiority of our ASD.
****
Neural Congealing: Aligning Images to a Joint Semantic Atlas
Dolev Ofri-Amar, Michal Geyer, Yoni Kasten, Tali Dekel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19403-19412
We present Neural Congealing -- a zero-shot self-supervised framework for detecting and jointly aligning semantically-common content across a given set of images. Our approach harnesses the power of pre-trained DINO-ViT features to learn: (i) a joint semantic atlas -- a 2D grid that captures the mode of DINO-ViT features in the input set, and (ii) dense mappings from the unified atlas to each of the input images. We derive a new robust self-supervised framework that optimizes the atlas representation and mappings per image set, requiring only a few real-world images as input without any additional input information (e.g., segmentation masks). Notably, we design our losses and training paradigm to account only for the shared content under severe variations in appearance, pose, background clutter or other distracting objects. We demonstrate results on a plethora of challenging image sets including sets of mixed domains (e.g., aligning images depicting sculpture and artwork of cats), sets depicting related yet different object categories (e.g., dogs and tigers), or domains for which large-scale training data is scarce (e.g., coffee mugs). We thoroughly evaluate our method and show that our test-time optimization approach performs favorably compared to a state-of-the-art method that requires extensive training on large-scale datasets.
****
Adaptive Spot-Guided Transformer for Consistent Local Feature Matching
Jiahuan Yu, Jiahao Chang, Jianfeng He, Tianzhu Zhang, Jiyang Yu, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21898-21908
Local feature matching aims at finding correspondences between a pair of images. Although current detector-free methods leverage Transformer architecture to obtain an impressive performance, few works consider maintaining local consistency. Meanwhile, most methods struggle with large scale variations. To deal with the above issues, we propose Adaptive Spot-Guided Transformer (ASTR) for local feature matching, which jointly models the local consistency and scale variations in a unified coarse-to-fine architecture. The proposed ASTR enjoys several merits. First, we design a spot-guided aggregation module to avoid interfering with irrelevant areas during feature aggregation. Second, we design an adaptive scaling module to adjust the size of grids according to the calculated depth information at fine stage. Extensive experimental results on five standard benchmarks demonstrate that our ASTR performs favorably against state-of-the-art methods.Our code will be released on https://astr2023.github.io.
****
Wide-Angle Rectification via Content-Aware Conformal Mapping
Qi Zhang, Hongdong Li, Qing Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17357-17365
Despite the proliferation of ultra wide-angle lenses on smartphone cameras, such lenses often come with severe image distortion (e.g. curved linear structure, unnaturally skewed faces). Most existing rectification methods adopt a global war

ping transformation to undistort the input wide-angle image, yet their performan ces are not entirely satisfactory, leaving many unwanted residue distortions unc orrected or at the sacrifice of the intended wide FoV (field-of-view). This pape r proposes a new method to tackle these challenges. Specifically, we derive a lo cally-adaptive polar-domain conformal mapping to rectify a wide-angle image. Par ameters of the mapping are found automatically by analyzing image contents via d eep neural networks. Experiments on large number of photos have confirmed the su perior performance of the proposed method compared with all available previous m ethods.
****
Towards Stable Human Pose Estimation via Cross-View Fusion and Foot Stabilizatio n
Li’an Zhuo, Jian Cao, Qi Wang, Bang Zhang, Liefeng Bo; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 650-6 59
Towards stable human pose estimation from monocular images, there remain two mai n dilemmas. On the one hand, the different perspectives, i.e., front view, side view, and top view, appear the inconsistent performances due to the depth ambigu ity. On the other hand, foot posture plays a significant role in complicated hum an pose estimation, i.e., dance and sports, and foot-ground interaction, but unf ortunately, it is omitted in most general approaches and datasets. In this paper , we first propose the Cross-View Fusion (CVF) module to catch up with better 3D intermediate representation and alleviate the view inconsistency based on the v ision transformer encoder. Then the optimization-based method is introduced to r econstruct the foot pose and foot-ground contact for the general multi-view data sets including AIST++ and Human3.6M. Besides, the reversible kinematic topology strategy is innovated to utilize the contact information into the full-body with foot pose regressor. Extensive experiments on the popular benchmarks demonstrat e that our method outperforms the state-of-the-art approaches by achieving 40.1m m PA-MPJPE on the 3DPW test set and 43.8mm on the AIST++ test set.
****
Few-Shot Non-Line-of-Sight Imaging With Signal-Surface Collaborative Regularizat ion
Xintong Liu, Jianyu Wang, Leping Xiao, Xing Fu, Lingyun Qiu, Zuoqiang Shi; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2023, pp. 13303-13312
The non-line-of-sight imaging technique aims to reconstruct targets from multipl y reflected light. For most existing methods, dense points on the relay surface are raster scanned to obtain high-quality reconstructions, which requires a long acquisition time. In this work, we propose a signal-surface collaborative regul arization (SSCR) framework that provides noise-robust reconstructions with a min imal number of measurements. Using Bayesian inference, we design joint regulariz ations of the estimated signal, the 3D voxel-based representation of the objects , and the 2D surface-based description of the targets. To our best knowledge, th is is the first work that combines regularizations in mixed dimensions for hidde n targets. Experiments on synthetic and experimental datasets illustrated the ef ficiency of the proposed method under both confocal and non-confocal settings. W e report the reconstruction of the hidden targets with complex geometric structu res with only 5 x 5 confocal measurements from public datasets, indicating an ac celeration of the conventional measurement process by a factor of 10,000. Beside s, the proposed method enjoys low time and memory complexity with sparse measure ments. Our approach has great potential in real-time non-line-of-sight imaging a pplications such as rescue operations and autonomous driving.
****
SINE: SINgle Image Editing With Text-to-Image Diffusion Models
Zhixing Zhang, Ligong Han, Arnab Ghosh, Dimitris N. Metaxas, Jian Ren; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6027-6037
Recent works on diffusion models have demonstrated a strong capability for condi tioning image generation, e.g., text-guided image synthesis. Such success inspir

es many efforts trying to use large-scale pre-trained diffusion models for tackling a challenging problem--real image editing. Works conducted in this area learn a unique textual token corresponding to several images containing the same object. However, under many circumstances, only one image is available, such as the painting of the Girl with a Pearl Earring. Using existing works on fine-tuning the pre-trained diffusion models with a single image causes severe overfitting issues. The information leakage from the pre-trained diffusion models makes editing can not keep the same content as the given image while creating new features depicted by the language guidance. This work aims to address the problem of single-image editing. We propose a novel model-based guidance built upon the classifier-free guidance so that the knowledge from the model trained on a single image can be distilled into the pre-trained diffusion model, enabling content creation even with one given image. Additionally, we propose a patch-based fine-tuning that can effectively help the model generate images of arbitrary resolution. We provide extensive experiments to validate the design choices of our approach and show promising editing capabilities, including changing style, content addition, and object manipulation.
****
Probabilistic Debiasing of Scene Graphs
Bashirul Azam Biswas, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10429-10438
The quality of scene graphs generated by the state-of-the-art (SOTA) models is compromised due to the long-tail nature of the relationships and their parent object pairs. Training of the scene graphs is dominated by the majority relationships of the majority pairs and, therefore, the object-conditional distributions of relationship in the minority pairs are not preserved after the training is converged. Consequently, the biased model performs well on more frequent relationships in the marginal distribution of relationships such as 'on' and 'wearing', and performs poorly on the less frequent relationships such as 'eating' or 'hanging from'. In this work, we propose virtual evidence incorporated within-triplet Bayesian Network (BN) to preserve the object-conditional distribution of the relationship label and to eradicate the bias created by the marginal probability of the relationships. The insufficient number of relationships in the minority classes poses a significant problem in learning the within-triplet Bayesian network. We address this insufficiency by embedding-based augmentation of triplets where we borrow samples of the minority triplet classes from its neighboring triplets in the semantic space. We perform experiments on two different datasets and achieve a significant improvement in the mean recall of the relationships. We also achieve a better balance between recall and mean recall performance compared to the SOTA de-biasing techniques of scene graph models.
****
OSAN: A One-Stage Alignment Network To Unify Multimodal Alignment and Unsupervised Domain Adaptation
Ye Liu, Lingfeng Qiao, Changchong Lu, Di Yin, Chen Lin, Haoyuan Peng, Bo Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3551-3560
Extending from unimodal to multimodal is a critical challenge for unsupervised domain adaptation (UDA). Two major problems emerge in unsupervised multimodal domain adaptation: domain adaptation and modality alignment. An intuitive way to handle these two problems is to fulfill these tasks in two separate stages: aligning modalities followed by domain adaptation, or vice versa. However, domains and modalities are not associated in most existing two-stage studies, and the relationship between them is not leveraged which can provide complementary information to each other. In this paper, we unify these two stages into one to align domains and modalities simultaneously. In our model, a tensor-based alignment module (TAL) is presented to explore the relationship between domains and modalities. By this means, domains and modalities can interact sufficiently and guide them to utilize complementary information for better results. Furthermore, to establish a bridge between domains, a dynamic domain generator (DDG) module is proposed to build transitional samples by mixing the shared information of two domains in

 a self-supervised manner, which helps our model learn a domain-invariant common
 representation space. Extensive experiments prove that our method can achieve s
uperior performance in two real-world applications. The code will be publicly av
ailable.
****
Token Turing Machines
Michael S. Ryoo, Keerthana Gopalakrishnan, Kumara Kahatapitiya, Ted Xiao, Kanish
ka Rao, Austin Stone, Yao Lu, Julian Ibarz, Anurag Arnab; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19
070-19081
We propose Token Turing Machines (TTM), a sequential, autoregressive Transformer
 model with memory for real-world sequential visual understanding. Our model is
inspired by the seminal Neural Turing Machine, and has an external memory consis
ting of a set of tokens which summarise the previous history (i.e., frames). Thi
s memory is efficiently addressed, read and written using a Transformer as the p
rocessing unit/controller at each step. The model's memory module ensures that a
 new observation will only be processed with the contents of the memory (and not
 the entire history), meaning that it can efficiently process long sequences wit
h a bounded computational cost at each step. We show that TTM outperforms other
alternatives, such as other Transformer models designed for long sequences and r
ecurrent neural networks, on two real-world sequential visual understanding task
s: online temporal activity detection from videos and vision-based robot action
policy learning. Code is publicly available at: https://github.com/google-resear
ch/scenic/tree/main/scenic/projects/token_turing.
****
Solving 3D Inverse Problems Using Pre-Trained 2D Diffusion Models
Hyungjin Chung, Dohoon Ryu, Michael T. McCann, Marc L. Klasky, Jong Chul Ye; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 22542-22551
Diffusion models have emerged as the new state-of-the-art generative model with
high quality samples, with intriguing properties such as mode coverage and high
flexibility. They have also been shown to be effective inverse problem solvers,
acting as the prior of the distribution, while the information of the forward mo
del can be granted at the sampling stage. Nonetheless, as the generative process
 remains in the same high dimensional (i.e. identical to data dimension) space,
the models have not been extended to 3D inverse problems due to the extremely hi
gh memory and computational cost. In this paper, we combine the ideas from the c
onventional model-based iterative reconstruction with the modern diffusion model
s, which leads to a highly effective method for solving 3D medical image reconst
ruction tasks such as sparse-view tomography, limited angle tomography, compress
ed sensing MRI from pre-trained 2D diffusion models. In essence, we propose to a
ugment the 2D diffusion prior with a model-based prior in the remaining directio
n at test time, such that one can achieve coherent reconstructions across all di
mensions. Our method can be run in a single commodity GPU, and establishes the n
ew state-of-the-art, showing that the proposed method can perform reconstruction
s of high fidelity and accuracy even in the most extreme cases (e.g. 2-view 3D t
omography). We further reveal that the generalization capacity of the proposed m
ethod is surprisingly high, and can be used to reconstruct volumes that are enti
rely different from the training dataset. Code available: https://github.com/HJ-
harry/DiffusionMBIR
****
Heat Diffusion Based Multi-Scale and Geometric Structure-Aware Transformer for M
esh Segmentation
Chi-Chong Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa
ttern Recognition (CVPR), 2023, pp. 4413-4422
Triangle mesh segmentation is an important task in 3D shape analysis, especially
 in applications such as digital humans and AR/VR. Transformer model is inherent
ly permutation-invariant to input, which makes it a suitable candidate model for
 3D mesh processing. However, two main challenges involved in adapting Transform
er from natural languages to 3D mesh are yet to be solved, such as i) extracting

the multi-scale information of mesh data in an adaptive manner; ii) capturing geometric structures of mesh data as the discriminative characteristics of the shape. Current point based Transformer models fail to tackle such challenges and thus provide inferior performance for discretized surface segmentation. In this work, heat diffusion based method is exploited to tackle these problems. A novel Transformer model called MeshFormer is proposed, which i) integrates Heat Diffusion method into Multi-head Self-Attention operation (HDMSA) to adaptively capture the features from local neighborhood to global contexts; ii) applies a novel Heat Kernel Signature based Structure Encoding (HKSSE) to embed the intrinsic geometric structures of mesh instances into Transformer for structure-aware processing. Extensive experiments on triangle mesh segmentation validate the effectiveness of the proposed MeshFormer model and show significant improvements over current state-of-the-art methods.
****
DyNCA: Real-Time Dynamic Texture Synthesis Using Neural Cellular Automata
Ehsan Pajouheshgar, Yitao Xu, Tong Zhang, Sabine Süsstrunk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20742-20751
Current Dynamic Texture Synthesis (DyTS) models can synthesize realistic videos. However, they require a slow iterative optimization process to synthesize a single fixed-size short video, and they do not offer any post-training control over the synthesis process. We propose Dynamic Neural Cellular Automata (DyNCA), a framework for real-time and controllable dynamic texture synthesis. Our method is built upon the recently introduced NCA models and can synthesize infinitely long and arbitrary-size realistic video textures in real-time. We quantitatively and qualitatively evaluate our model and show that our synthesized videos appear more realistic than the existing results. We improve the SOTA DyTS performance by 2 4 orders of magnitude. Moreover, our model offers several real-time video controls including motion speed, motion direction, and an editing brush tool. We exhibit our trained models in an online interactive demo that runs on local hardware and is accessible on personal computers and smartphones.
****
Semantic-Promoted Debiasing and Background Disambiguation for Zero-Shot Instance Segmentation
Shuting He, Henghui Ding, Wei Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19498-19507
Zero-shot instance segmentation aims to detect and precisely segment objects of unseen categories without any training samples. Since the model is trained on seen categories, there is a strong bias that the model tends to classify all the objects into seen categories. Besides, there is a natural confusion between background and novel objects that have never shown up in training. These two challenges make novel objects hard to be raised in the final instance segmentation results. It is desired to rescue novel objects from background and dominated seen categories. To this end, we propose D^2Zero with Semantic-Promoted Debiasing and Background Disambiguation to enhance the performance of Zero-shot instance segmentation. Semantic-promoted debiasing utilizes inter-class semantic relationships to involve unseen categories in visual feature training and learns an input-conditional classifier to conduct dynamical classification based on the input image. Background disambiguation produces image-adaptive background representation to avoid mistaking novel objects for background. Extensive experiments show that we significantly outperform previous state-of-the-art methods by a large margin, e.g., 16.86% improvement on COCO.
****
RelightableHands: Efficient Neural Relighting of Articulated Hand Models
Shun Iwase, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Timur Bagautdinov, Rohan Joshi, Fabian Prada, Takaaki Shiratori, Yaser Sheikh, Jason Saragih; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16663-16673
We present the first neural relighting approach for rendering high-fidelity personalized hands that can be animated in real-time under novel illumination. Our a

pproach adopts a teacher-student framework, where the teacher learns appearance under a single point light from images captured in a light-stage, allowing us to synthesize hands in arbitrary illuminations but with heavy compute. Using images rendered by the teacher model as training data, an efficient student model directly predicts appearance under natural illuminations in real-time. To achieve generalization, we condition the student model with physics-inspired illumination features such as visibility, diffuse shading, and specular reflections computed on a coarse proxy geometry, maintaining a small computational overhead. Our key insight is that these features have strong correlation with subsequent global light transport effects, which proves sufficient as conditioning data for the neural relighting network. Moreover, in contrast to bottleneck illumination conditioning, these features are spatially aligned based on underlying geometry, leading to better generalization to unseen illuminations and poses. In our experiments, we demonstrate the efficacy of our illumination feature representations, outperforming baseline approaches. We also show that our approach can photorealistically relight two interacting hands at real-time speeds.
****
Paired-Point Lifting for Enhanced Privacy-Preserving Visual Localization
Chunghwan Lee, Jaihoon Kim, Chanhyuk Yun, Je Hyeong Hong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17266-17275
Visual localization refers to the process of recovering camera pose from input image relative to a known scene, forming a cornerstone of numerous vision and robotics systems. While many algorithms utilize sparse 3D point cloud of the scene obtained via structure-from-motion (SfM) for localization, recent studies have raised privacy concerns by successfully revealing high-fidelity appearance of the scene from such sparse 3D representation. One prominent approach for bypassing this attack was to lift 3D points to randomly oriented 3D lines thereby hiding scene geometry, but latest work have shown such random line cloud has a critical statistical flaw that can be exploited to break through protection. In this work, we present an alternative lightweight strategy called Paired-Point Lifting (PPL) for constructing 3D line clouds. Instead of drawing one randomly oriented line per 3D point, PPL splits 3D points into pairs and joins each pair to form 3D lines. This seemingly simple strategy yields 3 benefits, i) new ambiguity in feature selection, ii) increased line cloud sparsity, and iii) non-trivial distribution of 3D lines, all of which contributes to enhanced protection against privacy attacks. Extensive experimental results demonstrate the strength of PPL in concealing scene details without compromising localization accuracy, unlocking the true potential of 3D line clouds.
****
Depth Estimation From Camera Image and mmWave Radar Point Cloud
Akash Deep Singh, Yunhao Ba, Ankur Sarker, Howard Zhang, Achuta Kadambi, Stefano Soatto, Mani Srivastava, Alex Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9275-9285
We present a method for inferring dense depth from a camera image and a sparse noisy radar point cloud. We first describe the mechanics behind mmWave radar point cloud formation and the challenges that it poses, i.e. ambiguous elevation and noisy depth and azimuth components that yields incorrect positions when projected onto the image, and how existing works have overlooked these nuances in camera-radar fusion. Our approach is motivated by these mechanics, leading to the design of a network that maps each radar point to the possible surfaces that it may project onto in the image plane. Unlike existing works, we do not process the raw radar point cloud as an erroneous depth map, but query each raw point independently to associate it with likely pixels in the image -- yielding a semi-dense radar depth map. To fuse radar depth with an image, we propose a gated fusion scheme that accounts for the confidence scores of the correspondence so that we selectively combine radar and camera embeddings to yield a dense depth map. We test our method on the NuScenes benchmark and show a 10.3% improvement in mean absolute error and a 9.1% improvement in root-mean-square error over the best method.

****
Learning Event Guided High Dynamic Range Video Reconstruction
Yixin Yang, Jin Han, Jinxiu Liang, Imari Sato, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13924-13934
Limited by the trade-off between frame rate and exposure time when capturing moving scenes with conventional cameras, frame based HDR video reconstruction suffers from scene-dependent exposure ratio balancing and ghosting artifacts. Event cameras provide an alternative visual representation with a much higher dynamic range and temporal resolution free from the above issues, which could be an effective guidance for HDR imaging from LDR videos. In this paper, we propose a multimodal learning framework for event guided HDR video reconstruction. In order to better leverage the knowledge of the same scene from the two modalities of visual signals, a multimodal representation alignment strategy to learn a shared latent space and a fusion module tailored to complementing two types of signals for different dynamic ranges in different regions are proposed. Temporal correlations are utilized recurrently to suppress the flickering effects in the reconstructed HDR video. The proposed HDRev-Net demonstrates state-of-the-art performance quantitatively and qualitatively for both synthetic and real-world data.
****
Multi-Granularity Archaeological Dating of Chinese Bronze Dings Based on a Knowledge-Guided Relation Graph
Rixin Zhou, Jiafu Wei, Qian Zhang, Ruihua Qi, Xi Yang, Chuntao Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3103-3113
The archaeological dating of bronze dings has played a critical role in the study of ancient Chinese history. Current archaeology depends on trained experts to carry out bronze dating, which is time-consuming and labor-intensive. For such dating, in this study, we propose a learning-based approach to integrate advanced deep learning techniques and archaeological knowledge. To achieve this, we first collect a large-scale image dataset of bronze dings, which contains richer attribute information than other existing fine-grained datasets. Second, we introduce a multihead classifier and a knowledge-guided relation graph to mine the relationship between attributes and the ding era. Third, we conduct comparison experiments with various existing methods, the results of which show that our dating method achieves a state-of-the-art performance. We hope that our data and applied networks will enrich fine-grained classification research relevant to other interdisciplinary areas of expertise. The dataset and source code used are included in our supplementary materials, and will be open after submission owing to the anonymity policy. Source codes and data are available at: https://github.com/zhourixin/bronze-Ding.
****
CASP-Net: Rethinking Video Saliency Prediction From an Audio-Visual Consistency Perceptual Perspective
Junwen Xiong, Ganglai Wang, Peng Zhang, Wei Huang, Yufei Zha, Guangtao Zhai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6441-6450
Incorporating the audio stream enables Video Saliency Prediction (VSP) to imitate the selective attention mechanism of human brain. By focusing on the benefits of joint auditory and visual information, most VSP methods are capable of exploiting semantic correlation between vision and audio modalities but ignoring the negative effects due to the temporal inconsistency of audio-visual intrinsics. Inspired by the biological inconsistency-correction within multi-sensory information, in this study, a consistency-aware audio-visual saliency prediction network (CASP-Net) is proposed, which takes a comprehensive consideration of the audio-visual semantic interaction and consistent perception. In addition a two-stream encoder for elegant association between video frames and corresponding sound source, a novel consistency-aware predictive coding is also designed to improve the consistency within audio and visual representations iteratively. To further aggregate the multi-scale audio-visual information, a saliency decoder is introduced

for the final saliency map generation. Substantial experiments demonstrate that the proposed CASP-Net outperforms the other state-of-the-art methods on six challenging audio-visual eye-tracking datasets. For a demo of our system please see https://woshihaozhu.github.io/CASP-Net/.
****
Learning Expressive Prompting With Residuals for Vision Transformers
Rajshekhar Das, Yonatan Dukler, Avinash Ravichandran, Ashwin Swaminathan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3366-3377
Prompt learning is an efficient approach to adapt transformers by inserting learnable set of parameters into the input and intermediate representations of a pre-trained model. In this work, we present Expressive Prompts with Residuals (EXPRES) which modifies the prompt learning paradigm specifically for effective adaptation of vision transformers (ViT). Out method constructs downstream representations via learnable "output" tokens, that are akin to the learned class tokens of the ViT. Further for better steering of the downstream representation processed by the frozen transformer, we introduce residual learnable tokens that are added to the output of various computations. We apply EXPRES for image classification, few shot learning, and semantic segmentation, and show our method is capable of achieving state of the art prompt tuning on 3/3 categories of the VTAB benchmark. In addition to strong performance, we observe that our approach is an order of magnitude more prompt efficient than existing visual prompting baselines. We analytically show the computational benefits of our approach over weight space adaptation techniques like finetuning. Lastly we systematically corroborate the architectural design of our method via a series of ablation experiments.
****
Prototypical Residual Networks for Anomaly Detection and Localization
Hui Zhang, Zuxuan Wu, Zheng Wang, Zhineng Chen, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16281-16291
Anomaly detection and localization are widely used in industrial manufacturing for its efficiency and effectiveness. Anomalies are rare and hard to collect and supervised models easily over-fit to these seen anomalies with a handful of abnormal samples, producing unsatisfactory performance. On the other hand, anomalies are typically subtle, hard to discern, and of various appearance, making it difficult to detect anomalies and let alone locate anomalous regions. To address these issues, we propose a framework called Prototypical Residual Network (PRN), which learns feature residuals of varying scales and sizes between anomalous and normal patterns to accurately reconstruct the segmentation maps of anomalous regions. PRN mainly consists of two parts: multi-scale prototypes that explicitly represent the residual features of anomalies to normal patterns; a multi-size self-attention mechanism that enables variable-sized anomalous feature learning. Besides, we present a variety of anomaly generation strategies that consider both seen and unseen appearance variance to enlarge and diversify anomalies. Extensive experiments on the challenging and widely used MVTec AD benchmark show that PRN outperforms current state-of-the-art unsupervised and supervised methods. We further report SOTA results on three additional datasets to demonstrate the effectiveness and generalizability of PRN.
****
What Happened 3 Seconds Ago? Inferring the Past With Thermal Imaging
Zitian Tang, Wenjie Ye, Wei-Chiu Ma, Hang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17111-17120
Inferring past human motion from RGB images is challenging due to the inherent uncertainty of the prediction problem. Thermal images, on the other hand, encode traces of past human-object interactions left in the environment via thermal radiation measurement. Based on this observation, we collect the first RGB-Thermal dataset for human motion analysis, dubbed Thermal-IM. Then we develop a three-stage neural network model for accurate past human pose estimation. Comprehensive experiments show that thermal cues significantly reduce the ambiguities of this task, and the proposed model achieves remarkable performance. The dataset is ava

ilable at https://github.com/ZitianTang/Thermal-IM.
****
Ultrahigh Resolution Image/Video Matting With Spatio-Temporal Sparsity
Yanan Sun, Chi-Keung Tang, Yu-Wing Tai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14112-14121
Commodity ultra-high definition (UHD) displays are becoming more affordable which demand imaging in ultra high resolution (UHR). This paper proposes SparseMat, a computationally efficient approach for UHR image/video matting. Note that it is infeasible to directly process UHR images at full resolution in one shot using existing matting algorithms without running out of memory on consumer-level computational platforms, e.g., Nvidia 1080Ti with 11G memory, while patch-based approaches can introduce unsightly artifacts due to patch partitioning. Instead, our method resorts to spatial and temporal sparsity for solving general UHR matting. During processing videos, huge computation redundancy can be reduced through the rational use of spatial and temporal sparsity. In this paper, we show how to effectively estimate spatio-temporal sparsity, which serves as a gate to activate input pixels for the matting model. Under the guidance of such sparsity, our method discards patch-based inference in lieu of memory-efficient and full-resolution matte refinement. Extensive experiments demonstrate that SparseMat can effectively and efficiently generate high-quality alpha matte for UHR images and videos in one shot.
****
AnyFlow: Arbitrary Scale Optical Flow With Implicit Neural Representation
Hyunyoung Jung, Zhuo Hui, Lei Luo, Haitao Yang, Feng Liu, Sungjoo Yoo, Rakesh Ranjan, Denis Demandolx; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5455-5465
To apply optical flow in practice, it is often necessary to resize the input to smaller dimensions in order to reduce computational costs. However, downsizing inputs makes the estimation more challenging because objects and motion ranges become smaller. Even though recent approaches have demonstrated high-quality flow estimation, they tend to fail to accurately model small objects and precise boundaries when the input resolution is lowered, restricting their applicability to high-resolution inputs. In this paper, we introduce AnyFlow, a robust network that estimates accurate flow from images of various resolutions. By representing optical flow as a continuous coordinate-based representation, AnyFlow generates outputs at arbitrary scales from low-resolution inputs, demonstrating superior performance over prior works in capturing tiny objects with detail preservation on a wide range of scenes. We establish a new state-of-the-art performance of cross-dataset generalization on the KITTI dataset, while achieving comparable accuracy on the online benchmarks to other SOTA methods.
****
Zero-Shot Noise2Noise: Efficient Image Denoising Without Any Data
Youssef Mansour, Reinhard Heckel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14018-14027
Recently, self-supervised neural networks have shown excellent image denoising performance. However, current dataset free methods are either computationally expensive, require a noise model, or have inadequate image quality. In this work we show that a simple 2-layer network, without any training data or knowledge of the noise distribution, can enable high-quality image denoising at low computational cost. Our approach is motivated by Noise2Noise and Neighbor2Neighbor and works well for denoising pixel-wise independent noise. Our experiments on artificial, real-world camera, and microscope noise show that our method termed ZS-N2N (Zero Shot Noise2Noise) often outperforms existing dataset-free methods at a reduced cost, making it suitable for use cases with scarce data availability and limited compute.
****
Vector Quantization With Self-Attention for Quality-Independent Representation Learning
Zhou Yang, Weisheng Dong, Xin Li, Mengluan Huang, Yulin Sun, Guangming Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C

VPR), 2023, pp. 24438-24448
Recently, the robustness of deep neural networks has drawn extensive attention d
ue to the potential distribution shift between training and testing data (e.g.,
deep models trained on high-quality images are sensitive to corruption during te
sting). Many researchers attempt to make the model learn invariant representatio
ns from multiple corrupted data through data augmentation or image-pair-based fe
ature distillation to improve the robustness. Inspired by sparse representation
in image restoration, we opt to address this issue by learning image-quality-ind
ependent feature representation in a simple plug-and-play manner, that is, to in
troduce discrete vector quantization (VQ) to remove redundancy in recognition mo
dels. Specifically, we first add a codebook module to the network to quantize de
ep features. Then we concatenate them and design a self-attention module to enha
nce the representation. During training, we enforce the quantization of features
 from clean and corrupted images in the same discrete embedding space so that an
 invariant quality-independent feature representation can be learned to improve
the recognition robustness of low-quality images. Qualitative and quantitative e
xperimental results show that our method achieved this goal effectively, leading
 to a new state-of-the-art result of 43.1% mCE on ImageNet-C with ResNet50 as th
e backbone. On other robustness benchmark datasets, such as ImageNet-R, our meth
od also has an accuracy improvement of almost 2%.
****
Generating Anomalies for Video Anomaly Detection With Prompt-Based Feature Mappi
ng
Zuhao Liu, Xiao-Ming Wu, Dian Zheng, Kun-Yu Lin, Wei-Shi Zheng; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 24500-24510
Anomaly detection in surveillance videos is a challenging computer vision task w
here only normal videos are available during training. Recent work released the
first virtual anomaly detection dataset to assist real-world detection. However,
 an anomaly gap exists because the anomalies are bounded in the virtual dataset
but unbounded in the real world, so it reduces the generalization ability of the
 virtual dataset. There also exists a scene gap between virtual and real scenari
os, including scene-specific anomalies (events that are abnormal in one scene bu
t normal in another) and scene-specific attributes, such as the viewpoint of the
 surveillance camera. In this paper, we aim to solve the problem of the anomaly
gap and scene gap by proposing a prompt-based feature mapping framework (PFMF).
The PFMF contains a mapping network guided by an anomaly prompt to generate unse
en anomalies with unbounded types in the real scenario, and a mapping adaptation
 branch to narrow the scene gap by applying domain classifier and anomaly classi
fier. The proposed framework outperforms the state-of-the-art on three benchmark
 datasets. Extensive ablation experiments also show the effectiveness of our fra
mework design.
****
Diffusion-Based Signed Distance Fields for 3D Shape Generation
Jaehyeok Shim, Changwoo Kang, Kyungdon Joo; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20887-20897
We propose a 3D shape generation framework (SDF-Diffusion in short) that uses de
noising diffusion models with continuous 3D representation via signed distance f
ields (SDF). Unlike most existing methods that depend on discontinuous forms, su
ch as point clouds, SDF-Diffusion generates high-resolution 3D shapes while alle
viating memory issues by separating the generative process into two-stage: gener
ation and super-resolution. In the first stage, a diffusion-based generative mod
el generates a low-resolution SDF of 3D shapes. Using the estimated low-resoluti
on SDF as a condition, the second stage diffusion model performs super-resolutio
n to generate high-resolution SDF. Our framework can generate a high-fidelity 3D
 shape despite the extreme spatial complexity. On the ShapeNet dataset, our mode
l shows competitive performance to the state-of-the-art methods and shows applic
ability on the shape completion task without modification.
****
Hierarchical Temporal Transformer for 3D Hand Pose Estimation and Action Recogni

tion From Egocentric RGB Videos
Yilin Wen, Hao Pan, Lei Yang, Jia Pan, Taku Komura, Wenping Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21243-21253
Understanding dynamic hand motions and actions from egocentric RGB videos is a fundamental yet challenging task due to self-occlusion and ambiguity. To address occlusion and ambiguity, we develop a transformer-based framework to exploit temporal information for robust estimation. Noticing the different temporal granularity of and the semantic correlation between hand pose estimation and action recognition, we build a network hierarchy with two cascaded transformer encoders, where the first one exploits the short-term temporal cue for hand pose estimation, and the latter aggregates per-frame pose and object information over a longer time span to recognize the action. Our approach achieves competitive results on two first-person hand action benchmarks, namely FPHA and H2O. Extensive ablation studies verify our design choices.
****
CAP-VSTNet: Content Affinity Preserved Versatile Style Transfer
Linfeng Wen, Chengying Gao, Changqing Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18300-18309
Content affinity loss including feature and pixel affinity is a main problem which leads to artifacts in photorealistic and video style transfer. This paper proposes a new framework named CAP-VSTNet, which consists of a new reversible residual network and an unbiased linear transform module, for versatile style transfer. This reversible residual network can not only preserve content affinity but not introduce redundant information as traditional reversible networks, and hence facilitate better stylization. Empowered by Matting Laplacian training loss which can address the pixel affinity loss problem led by the linear transform, the proposed framework is applicable and effective on versatile style transfer. Extensive experiments show that CAP-VSTNet can produce better qualitative and quantitative results in comparison with the state-of-the-art methods.
****
FIANCEE: Faster Inference of Adversarial Networks via Conditional Early Exits
Polina Karpikova, Ekaterina Radionova, Anastasia Yaschenko, Andrei Spiridonov, Leonid Kostyushko, Riccardo Fabbricatore, Aleksei Ivakhnenko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12032-12043
Generative DNNs are a powerful tool for image synthesis, but they are limited by their computational load. On the other hand, given a trained model and a task, e.g. faces generation within a range of characteristics, the output image quality will be unevenly distributed among images with different characteristics. It follows, that we might restrain the model's complexity on some instances, maintaining a high quality. We propose a method for diminishing computations by adding so-called early exit branches to the original architecture, and dynamically switching the computational path depending on how difficult it will be to render the output. We apply our method on two different SOTA models performing generative tasks: generation from a semantic map, and cross reenactment of face expressions; showing it is able to output images with custom lower quality thresholds. For a threshold of LPIPS <=0.1, we diminish their computations by up to a half. This is especially relevant for real-time applications such as synthesis of faces, when quality loss needs to be contained, but most of the inputs need fewer computations than the complex instances.
****
Simultaneously Short- and Long-Term Temporal Modeling for Semi-Supervised Video Semantic Segmentation
Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, Wei Chu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14763-14772
In order to tackle video semantic segmentation task at a lower cost, e.g., only one frame annotated per video, lots of efforts have been devoted to investigate the utilization of those unlabeled frames by either assigning pseudo labels or p

erforming feature enhancement. In this work, we propose a novel feature enhancement network to simultaneously model short- and long-term temporal correlation. Compared with existing work that only leverage short-term correspondence, the long-term temporal correlation obtained from distant frames can effectively expand the temporal perception field and provide richer contextual prior. More importantly, modeling adjacent and distant frames together can alleviate the risk of over-fitting, hence produce high-quality feature representation for the distant unlabeled frames in training set and unseen videos in testing set. To this end, we term our method SSLTM, short for Simultaneously Short- and Long-Term Temporal Modeling. In the setting of only one frame annotated per video, SSLTM significantly outperforms the state-of-the-art methods by 2%   3% mIoU on the challenging VSPW dataset. Furthermore, when working with a pseudo label based method such as MeanTeacher, our final model only exhibits 0.13% mIoU less than the ceiling performance (i.e., all frames are manually annotated).
****
Federated Domain Generalization With Generalization Adjustment
Ruipeng Zhang, Qinwei Xu, Jiangchao Yao, Ya Zhang, Qi Tian, Yanfeng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3954-3963
Federated Domain Generalization (FedDG) attempts to learn a global model in a privacy-preserving manner that generalizes well to new clients possibly with domain shift. Recent exploration mainly focuses on designing an unbiased training strategy within each individual domain. However, without the support of multi-domain data jointly in the mini-batch training, almost all methods cannot guarantee the generalization under domain shift. To overcome this problem, we propose a novel global objective incorporating a new variance reduction regularizer to encourage fairness. A novel FL-friendly method named Generalization Adjustment (GA) is proposed to optimize the above objective by dynamically calibrating the aggregation weights. The theoretical analysis of GA demonstrates the possibility to achieve a tighter generalization bound with an explicit re-weighted aggregation, substituting the implicit multi-domain data sharing that is only applicable to the conventional DG settings. Besides, the proposed algorithm is generic and can be combined with any local client training-based methods. Extensive experiments on several benchmark datasets have shown the effectiveness of the proposed method, with consistent improvements over several FedDG algorithms when used in combination. The source code is released at https://github.com/MediaBrain-SJTU/FedDG-GA.
****
Tunable Convolutions With Parametric Multi-Loss Optimization
Matteo Maggioni, Thomas Tanay, Francesca Babiloni, Steven McDonagh, Aleš Leonardis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20226-20236
Behavior of neural networks is irremediably determined by the specific loss and data used during training. However it is often desirable to tune the model at inference time based on external factors such as preferences of the user or dynamic characteristics of the data. This is especially important to balance the perception-distortion trade-off of ill-posed image-to-image translation tasks. In this work, we propose to optimize a parametric tunable convolutional layer, which includes a number of different kernels, using a parametric multi-loss, which includes an equal number of objectives. Our key insight is to use a shared set of parameters to dynamically interpolate both the objectives and the kernels. During training, these parameters are sampled at random to explicitly optimize all possible combinations of objectives and consequently disentangle their effect into the corresponding kernels. During inference, these parameters become interactive inputs of the model hence enabling reliable and consistent control over the model behavior. Extensive experimental results demonstrate that our tunable convolutions effectively work as a drop-in replacement for traditional convolutions in existing neural networks at virtually no extra computational cost, outperforming state-of-the-art control strategies in a wide range of applications; including image denoising, deblurring, super-resolution, and style transfer.

****
Learning To Generate Text-Grounded Mask for Open-World Semantic Segmentation Fro
m Only Image-Text Pairs
Junbum Cha, Jonghwan Mun, Byungseok Roh; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11165-11174
We tackle open-world semantic segmentation, which aims at learning to segment ar
bitrary visual concepts in images, by using only image-text pairs without dense
annotations. Existing open-world segmentation methods have shown impressive adva
nces by employing contrastive learning (CL) to learn diverse visual concepts and
 transferring the learned image-level understanding to the segmentation task. Ho
wever, these CL-based methods suffer from a train-test discrepancy, since it onl
y considers image-text alignment during training, whereas segmentation requires
region-text alignment during testing. In this paper, we proposed a novel Text-gr
ounded Contrastive Learning (TCL) framework that enables a model to directly lea
rn region-text alignment. Our method generates a segmentation mask for a given t
ext, extracts text-grounded image embedding from the masked region, and aligns i
t with text embedding via TCL. By learning region-text alignment directly, our f
ramework encourages a model to directly improve the quality of generated segment
ation masks. In addition, for a rigorous and fair comparison, we present a unifi
ed evaluation protocol with widely used 8 semantic segmentation datasets. TCL ac
hieves state-of-the-art zero-shot segmentation performances with large margins i
n all datasets. Code is available at https://github.com/kakaobrain/tcl.
****
CoMFormer: Continual Learning in Semantic and Panoptic Segmentation
Fabio Cermelli, Matthieu Cord, Arthur Douillard; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3010-3020
Continual learning for segmentation has recently seen increasing interest. Howev
er, all previous works focus on narrow semantic segmentation and disregard panop
tic segmentation, an important task with real-world impacts. In this paper, we p
resent the first continual learning model capable of operating on both semantic
and panoptic segmentation. Inspired by recent transformer approaches that consid
er segmentation as a mask-classification problem, we design CoMFormer. Our metho
d carefully exploits the properties of transformer architectures to learn new cl
asses over time. Specifically, we propose a novel adaptive distillation loss alo
ng with a mask-based pseudo-labeling technique to effectively prevent forgetting
. To evaluate our approach, we introduce a novel continual panoptic segmentation
 benchmark on the challenging ADE20K dataset. Our CoMFormer outperforms all the
existing baselines by forgetting less old classes but also learning more effecti
vely new classes. In addition, we also report an extensive evaluation in the lar
ge-scale continual semantic segmentation scenario showing that CoMFormer also si
gnificantly outperforms state-of-the-art methods.
****
DeepSolo: Let Transformer Decoder With Explicit Points Solo for Text Spotting
Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Tongliang Liu, Bo Du, Dacheng
Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2023, pp. 19348-19357
End-to-end text spotting aims to integrate scene text detection and recognition
into a unified framework. Dealing with the relationship between the two sub-task
s plays a pivotal role in designing effective spotters. Although Transformer-bas
ed methods eliminate the heuristic post-processing, they still suffer from the s
ynergy issue between the sub-tasks and low training efficiency. In this paper, w
e present DeepSolo, a simple DETR-like baseline that lets a single Decoder with
Explicit Points Solo for text detection and recognition simultaneously. Technica
lly, for each text instance, we represent the character sequence as ordered poin
ts and model them with learnable explicit point queries. After passing a single
decoder, the point queries have encoded requisite text semantics and locations,
thus can be further decoded to the center line, boundary, script, and confidence
 of text via very simple prediction heads in parallel. Besides, we also introduc
e a text-matching criterion to deliver more accurate supervisory signals, thus e
nabling more efficient training. Quantitative experiments on public benchmarks d

emonstrate that DeepSolo outperforms previous state-of-the-art methods and achieves better training efficiency. In addition, DeepSolo is also compatible with line annotations, which require much less annotation cost than polygons. The code is available at https://github.com/ViTAE-Transformer/DeepSolo.
****
Conditional Generation of Audio From Video via Foley Analogies
Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, Andrew Owens; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2426-2436
The sound effects that designers add to videos are designed to convey a particular artistic effect and, thus, may be quite different from a scene's true sound. Inspired by the challenges of creating a soundtrack for a video that differs from its true sound, but that nonetheless matches the actions occurring on screen, we propose the problem of conditional Foley. We present the following contributions to address this problem. First, we propose a pretext task for training our model to predict sound for an input video clip using a conditional audio-visual clip sampled from another time within the same source video. Second, we propose a model for generating a soundtrack for a silent input video, given a user-supplied example that specifies what the video should "sound like". We show through human studies and automated evaluation metrics that our model successfully generates sound from video, while varying its output according to the content of a supplied example.
****
Diverse 3D Hand Gesture Prediction From Body Dynamics by Bilateral Hand Disentanglement
Xingqun Qi, Chen Liu, Muyi Sun, Lincheng Li, Changjie Fan, Xin Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4616-4626
Predicting natural and diverse 3D hand gestures from the upper body dynamics is a practical yet challenging task in virtual avatar creation. Previous works usually overlook the asymmetric motions between two hands and generate two hands in a holistic manner, leading to unnatural results. In this work, we introduce a novel bilateral hand disentanglement based two-stage 3D hand generation method to achieve natural and diverse 3D hand prediction from body dynamics. In the first stage, we intend to generate natural hand gestures by two hand-disentanglement branches. Considering the asymmetric gestures and motions of two hands, we introduce a Spatial-Residual Memory (SRM) module to model spatial interaction between the body and each hand by residual learning. To enhance the coordination of two hand motions wrt. body dynamics holistically, we then present a Temporal-Motion Memory (TMM) module. TMM can effectively model the temporal association between body dynamics and two hand motions. The second stage is built upon the insight that 3D hand predictions should be non-deterministic given the sequential body postures. Thus, we further diversify our 3D hand predictions based on the initial output from the stage one. Concretely, we propose a Prototypical-Memory Sampling Strategy (PSS) to generate the non-deterministic hand gestures by gradient-based Markov Chain Monte Carlo (MCMC) sampling. Extensive experiments demonstrate that our method outperforms the state-of-the-art models on the B2H dataset and our newly collected TED Hands dataset. The dataset and code are available at: https://github.com/XingqunQi-lab/Diverse-3D-Hand-Gesture-Prediction.
****
DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation
Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22500-22510
Large text-to-image models achieved a remarkable leap in the evolution of AI, enabling high-quality and diverse synthesis of images from a given text prompt. However, these models lack the ability to mimic the appearance of subjects in a given reference set and synthesize novel renditions of them in different contexts. In this work, we present a new approach for "personalization" of text-to-image

diffusion models. Given as input just a few images of a subject, we fine-tune a pretrained text-to-image model such that it learns to bind a unique identifier with that specific subject. Once the subject is embedded in the output domain of the model, the unique identifier can be used to synthesize novel photorealistic images of the subject contextualized in different scenes. By leveraging the semantic prior embedded in the model with a new autogenous class-specific prior preservation loss, our technique enables synthesizing the subject in diverse scenes, poses, views and lighting conditions that do not appear in the reference images. We apply our technique to several previously-unassailable tasks, including subject recontextualization, text-guided view synthesis, and artistic rendering, all while preserving the subject's key features. We also provide a new dataset and evaluation protocol for this new task of subject-driven generation. Project page: https://dreambooth.github.io/
****
MOSO: Decomposing MOtion, Scene and Object for Video Prediction
Mingzhen Sun, Weining Wang, Xinxin Zhu, Jing Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18727-18737
Motion, scene and object are three primary visual components of a video. In particular, objects represent the foreground, scenes represent the background, and motion traces their dynamics. Based on this insight, we propose a two-stage MOtion, Scene and Object decomposition framework (MOSO) for video prediction, consisting of MOSO-VQVAE and MOSO-Transformer. In the first stage, MOSO-VQVAE decomposes a previous video clip into the motion, scene and object components, and represents them as distinct groups of discrete tokens. Then, in the second stage, MOSO-Transformer predicts the object and scene tokens of the subsequent video clip based on the previous tokens and adds dynamic motion at the token level to the generated object and scene tokens. Our framework can be easily extended to unconditional video generation and video frame interpolation tasks. Experimental results demonstrate that our method achieves new state-of-the-art performance on five challenging benchmarks for video prediction and unconditional video generation: BAIR, RoboNet, KTH, KITTI and UCF101. In addition, MOSO can produce realistic videos by combining objects and scenes from different videos.
****
Shakes on a Plane: Unsupervised Depth Estimation From Unstabilized Photography
Ilya Chugunov, Yuxuan Zhang, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13240-13251
Modern mobile burst photography pipelines capture and merge a short sequence of frames to recover an enhanced image, but often disregard the 3D nature of the scene they capture, treating pixel motion between images as a 2D aggregation problem. We show that in a "long-burst", forty-two 12-megapixel RAW frames captured in a two-second sequence, there is enough parallax information from natural hand tremor alone to recover high-quality scene depth. To this end, we devise a test-time optimization approach that fits a neural RGB-D representation to long-burst data and simultaneously estimates scene depth and camera motion. Our plane plus depth model is trained end-to-end, and performs coarse-to-fine refinement by controlling which multi-resolution volume features the network has access to at what time during training. We validate the method experimentally, and demonstrate geometrically accurate depth reconstructions with no additional hardware or separate data pre-processing and pose-estimation steps.
****
Learning Video Representations From Large Language Models
Yue Zhao, Ishan Misra, Philipp Krähenbühl, Rohit Girdhar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6586-6597
We introduce LAVILA, a new approach to learning video-language representations by leveraging Large Language Models (LLMs). We repurpose pre-trained LLMs to be conditioned on visual input, and finetune them to create automatic video narrators. Our auto-generated narrations offer a number of advantages, including dense coverage of long videos, better temporal synchronization of the visual informatio

n and text, and much higher diversity of text. The video-language embedding lear ned contrastively with these narrations outperforms the previous state-of-the-ar t on multiple first-person and third-person video tasks, both in zero-shot and f inetuned setups. Most notably, LAVILA obtains an absolute gain of 10.1% on EGTEA classification and 5.9% Epic-Kitchens-100 multi-instance retrieval benchmarks. Furthermore, LAVILA trained with only half the narrations from the Ego4D dataset outperforms models trained on the full set, and shows positive scaling behavior on increasing pre-training data and model size.
****
Learning the Distribution of Errors in Stereo Matching for Joint Disparity and U ncertainty Estimation
Liyan Chen, Weihan Wang, Philippos Mordohai; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17235-17244
We present a new loss function for joint disparity and uncertainty estimation in deep stereo matching. Our work is motivated by the need for precise uncertainty estimates and the observation that multi-task learning often leads to improved performance in all tasks. We show that this can be achieved by requiring the dis tribution of uncertainty to match the distribution of disparity errors via a KL divergence term in the network's loss function. A differentiable soft-histogramm ing technique is used to approximate the distributions so that they can be used in the loss. We experimentally assess the effectiveness of our approach and obse rve significant improvements in both disparity and uncertainty prediction on lar ge datasets. Our code is available at https://github.com/lly00412/SEDNet.git.
****
Learning Correspondence Uncertainty via Differentiable Nonlinear Least Squares
Dominik Muhle, Lukas Koestler, Krishna Murthy Jatavallabhula, Daniel Cremers; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13102-13112
We propose a differentiable nonlinear least squares framework to account for unc ertainty in relative pose estimation from feature correspondences. Specifically, we introduce a symmetric version of the probabilistic normal epipolar constrain t, and an approach to estimate the covariance of feature positions by differenti ating through the camera pose estimation procedure. We evaluate our approach on synthetic, as well as the KITTI and EuRoC real-world datasets. On the synthetic dataset, we confirm that our learned covariances accurately approximate the true noise distribution. In real world experiments, we find that our approach consis tently outperforms state-of-the-art non-probabilistic and probabilistic approach es, regardless of the feature extraction algorithm of choice.
****
Samples With Low Loss Curvature Improve Data Efficiency
Isha Garg, Kaushik Roy; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2023, pp. 20290-20300
In this paper, we study the second order properties of the loss of trained deep neural networks with respect to the training data points to understand the curva ture of the loss surface in the vicinity of these points. We find that there is an unexpected concentration of samples with very low curvature. We note that the se low curvature samples are largely consistent across completely different arch itectures, and identifiable in the early epochs of training. We show that the cu rvature relates to the 'cleanliness' of the data points, with low curvatures sam ples corresponding to clean, higher clarity samples, representative of their cat egory. Alternatively, high curvature samples are often occluded, have conflictin g features and visually atypical of their category. Armed with this insight, we introduce SLo-Curves, a novel coreset identification and training algorithm. SLo -curves identifies the samples with low curvatures as being more data-efficient and trains on them with an additional regularizer that penalizes high curvature of the loss surface in their vicinity. We demonstrate the efficacy of SLo-Curves on CIFAR-10 and CIFAR-100 datasets, where it outperforms state of the art cores et selection methods at small coreset sizes by up to 9%. The identified coresets generalize across architectures, and hence can be pre-computed to generate cond ensed versions of datasets for use in downstream tasks.

****
Towards Effective Visual Representations for Partial-Label Learning
Shiyu Xia, Jiaqi Lv, Ning Xu, Gang Niu, Xin Geng; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15589-1559
8
Under partial-label learning (PLL) where, for each training instance, only a set
 of ambiguous candidate labels containing the unknown true label is accessible,
contrastive learning has recently boosted the performance of PLL on vision tasks
, attributed to representations learned by contrasting the same/different classe
s of entities. Without access to true labels, positive points are predicted usin
g pseudolabels that are inherently noisy, and negative points often require larg
e batches or momentum encoders, resulting in unreliable similarity information a
nd a high computational overhead. In this paper, we rethink a state-of-the-art c
ontrastive PLL method PiCO [24], inspiring the design of a simple framework term
ed PaPi (Partial-label learning with a guided Prototypical classifier), which de
monstrates significant scope for improvement in representation learning, thus co
ntributing to label disambiguation. PaPi guides the optimization of a prototypic
al classifier by a linear classifier with which they share the same feature enco
der, thus explicitly encouraging the representation to reflect visual similarity
 between categories. It is also technically appealing, as PaPi requires only a f
ew components in PiCO with the opposite direction of guidance, and directly elim
inates the contrastive learning module that would introduce noise and consume co
mputational resources. We empirically demonstrate that PaPi significantly outper
forms other PLL methods on various image classification tasks.
****
MaskCLIP: Masked Self-Distillation Advances Contrastive Language-Image Pretraini
ng
Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Mi
ng Zeng, Weiming Zhang, Lu Yuan, Dong Chen, Fang Wen, Nenghai Yu; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 10995-11005
This paper presents a simple yet effective framework MaskCLIP, which incorporate
s a newly proposed masked self-distillation into contrastive language-image pret
raining. The core idea of masked self-distillation is to distill representation
from a full image to the representation predicted from a masked image. Such inco
rporation enjoys two vital benefits. First, masked self-distillation targets loc
al patch representation learning, which is complementary to vision-language cont
rastive focusing on text-related representation. Second, masked self-distillatio
n is also consistent with vision-language contrastive from the perspective of tr
aining objective as both utilize the visual encoder for feature aligning, and th
us is able to learn local semantics getting indirect supervision from the langua
ge. We provide specially designed experiments with a comprehensive analysis to v
alidate the two benefits. Symmetrically, we also introduce the local semantic su
pervision into the text branch, which further improves the pretraining performan
ce. With extensive experiments, we show that MaskCLIP, when applied to various c
hallenging downstream tasks, achieves superior results in linear probing, finetu
ning, and zero-shot performance with the guidance of the language encoder. We wi
ll release the code and data after the publication.
****
Open-Vocabulary Semantic Segmentation With Mask-Adapted CLIP
Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizha
o Zhang, Peter Vajda, Diana Marculescu; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7061-7070
Open-vocabulary semantic segmentation aims to segment an image into semantic reg
ions according to text descriptions, which may not have been seen during trainin
g. Recent two-stage methods first generate class-agnostic mask proposals and the
n leverage pre-trained vision-language models, e.g., CLIP, to classify masked re
gions. We identify the performance bottleneck of this paradigm to be the pre-tra
ined CLIP model, since it does not perform well on masked images. To address thi
s, we propose to finetune CLIP on a collection of masked image regions and their

 corresponding text descriptions. We collect training data by mining an existing
 image-caption dataset (e.g., COCO Captions), using CLIP to match masked image r
egions to nouns in the image captions. Compared with the more precise and manual
ly annotated segmentation labels with fixed classes (e.g., COCO-Stuff), we find
our noisy but diverse dataset can better retain CLIP's generalization ability. A
long with finetuning the entire model, we utilize the "blank" areas in masked im
ages using a method we dub mask prompt tuning. Experiments demonstrate mask prom
pt tuning brings significant improvement without modifying any weights of CLIP,
and it can further improve a fully finetuned model. In particular, when trained
on COCO and evaluated on ADE20K-150, our best model achieves 29.6% mIoU, which i
s +8.5% higher than the previous state-of-the-art. For the first time, open-voca
bulary generalist models match the performance of supervised specialist models i
n 2017 without dataset-specific adaptations.
****
A Loopback Network for Explainable Microvascular Invasion Classification
Shengxuming Zhang, Tianqi Shi, Yang Jiang, Xiuming Zhang, Jie Lei, Zunlei Feng,
Mingli Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2023, pp. 7443-7453
Microvascular invasion (MVI) is a critical factor for prognosis evaluation and c
ancer treatment. The current diagnosis of MVI relies on pathologists to manually
 find out cancerous cells from hundreds of blood vessels, which is time-consumin
g, tedious, and subjective. Recently, deep learning has achieved promising resul
ts in medical image analysis tasks. However, the unexplainability of black box m
odels and the requirement of massive annotated samples limit the clinical applic
ation of deep learning based diagnostic methods. In this paper, aiming to develo
p an accurate, objective, and explainable diagnosis tool for MVI, we propose a L
oopback Network (LoopNet) for classifying MVI efficiently. With the image-level
category annotations of the collected Pathologic Vessel Image Dataset (PVID), Lo
opNet is devised to be composed binary classification branch and cell locating b
ranch. The latter is devised to locate the area of cancerous cells, regular non-
cancerous cells, and background. For healthy samples, the pseudo masks of cells
supervise the cell locating branch to distinguish the area of regular non-cancer
ous cells and background. For each MVI sample, the cell locating branch predicts
 the mask of cancerous cells. Then the masked cancerous and non-cancerous areas
of the same sample are inputted back to the binary classification branch separat
ely. The loopback between two branches enables the category label to supervise t
he cell locating branch to learn the locating ability for cancerous areas. Exper
iment results show that the proposed LoopNet achieves 97.5% accuracy on MVI clas
sification. Surprisingly, the proposed loopback mechanism not only enables LoopN
et to predict the cancerous area but also facilitates the classification backbon
e to achieve better classification performance.
****
TINC: Tree-Structured Implicit Neural Compression
Runzhao Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 18517-18526
Implicit neural representation (INR) can describe the target scenes with high fi
delity using a small number of parameters, and is emerging as a promising data c
ompression technique. However, limited spectrum coverage is intrinsic to INR, an
d it is non-trivial to remove redundancy in diverse complex data effectively. Pr
eliminary studies can only exploit either global or local correlation in the tar
get data and thus of limited performance. In this paper, we propose a Tree-struc
tured Implicit Neural Compression (TINC) to conduct compact representation for l
ocal regions and extract the shared features of these local representations in a
 hierarchical manner. Specifically, we use Multi-Layer Perceptrons (MLPs) to fit
 the partitioned local regions, and these MLPs are organized in tree structure t
o share parameters according to the spatial distance. The parameter sharing sche
me not only ensures the continuity between adjacent regions, but also jointly re
moves the local and non-local redundancy. Extensive experiments show that TINC i
mproves the compression fidelity of INR, and has shown impressive compression ca
pabilities over commercial tools and other deep learning based methods. Besides,

 the approach is of high flexibility and can be tailored for different data and parameter settings. The source code can be found at https://github.com/RichealYoung/TINC.
****
Unifying Short and Long-Term Tracking With Graph Hierarchies
Orcun Cetintas, Guillem Brasó, Laura Leal-Taixé; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22877-22887
Tracking objects over long videos effectively means solving a spectrum of problems, from short-term association for un-occluded objects to long-term association for objects that are occluded and then reappear in the scene. Methods tackling these two tasks are often disjoint and crafted for specific scenarios, and top-performing approaches are often a mix of techniques, which yields engineering-heavy solutions that lack generality. In this work, we question the need for hybrid approaches and introduce SUSHI, a unified and scalable multi-object tracker. Our approach processes long clips by splitting them into a hierarchy of subclips, which enables high scalability. We leverage graph neural networks to process all levels of the hierarchy, which makes our model unified across temporal scales and highly general. As a result, we obtain significant improvements over state-of-the-art on four diverse datasets. Our code and models are available at bit.ly/sushi-mot.
****
Inferring and Leveraging Parts From Object Shape for Improving Semantic Image Synthesis
Yuxiang Wei, Zhilong Ji, Xiaohe Wu, Jinfeng Bai, Lei Zhang, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11248-11258
Despite the progress in semantic image synthesis, it remains a challenging problem to generate photo-realistic parts from input semantic map. Integrating part segmentation map can undoubtedly benefit image synthesis, but is bothersome and inconvenient to be provided by users. To improve part synthesis, this paper presents to infer Parts from Object ShapE (iPOSE) and leverage it for improving semantic image synthesis. However, albeit several part segmentation datasets are available, part annotations are still not provided for many object categories in semantic image synthesis. To circumvent it, we resort to few-shot regime to learn a PartNet for predicting the object part map with the guidance of pre-defined support part maps. PartNet can be readily generalized to handle a new object category when a small number (e.g., 3) of support part maps for this category are provided. Furthermore, part semantic modulation is presented to incorporate both inferred part map and semantic map for image synthesis. Experiments show that our iPOSE not only generates objects with rich part details, but also enables to control the image synthesis flexibly. And our iPOSE performs favorably against the state-of-the-art methods in terms of quantitative and qualitative evaluation. Our code will be publicly available at https://github.com/csyxwei/iPOSE.
****
MIME: Human-Aware 3D Scene Generation
Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, Michael J. Black; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12965-12976
Generating realistic 3D worlds occupied by moving humans has many applications in games, architecture, and synthetic data creation. But generating such scenes is expensive and labor intensive. Recent work generates human poses and motions given a 3D scene. Here, we take the opposite approach and generate 3D indoor scenes given 3D human motion. Such motions can come from archival motion capture or from IMU sensors worn on the body, effectively turning human movement in a "scanner" of the 3D world. Intuitively, human movement indicates the free-space in a room and human contact indicates surfaces or objects that support activities such as sitting, lying or touching. We propose MIME (Mining Interaction and Movement to infer 3D Environments), which is a generative model of indoor scenes that produces furniture layouts that are consistent with the human movement. MIME uses an auto-regressive transformer architecture that takes the already generated ob

jects in the scene as well as the human motion as input, and outputs the next plausible object. To train MIME, we build a dataset by populating the 3D FRONT scene dataset with 3D humans. Our experiments show that MIME produces more diverse and plausible 3D scenes than a recent generative scene method that does not know about human movement. Code and data will be available for research.
****
Re-Basin via Implicit Sinkhorn Differentiation
Fidel A. Guerrero Peña, Heitor Rapela Medeiros, Thomas Dubail, Masih Aminbeidokhti, Eric Granger, Marco Pedersoli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20237-20246
The recent emergence of new algorithms for permuting models into functionally equivalent regions of the solution space has shed some light on the complexity of error surfaces and some promising properties like mode connectivity. However, finding the permutation that minimizes some objectives is challenging, and current optimization techniques are not differentiable, which makes it difficult to integrate into a gradient-based optimization, and often leads to sub-optimal solutions. In this paper, we propose a Sinkhorn re-basin network with the ability to obtain the transportation plan that better suits a given objective. Unlike the current state-of-art, our method is differentiable and, therefore, easy to adapt to any task within the deep learning domain. Furthermore, we show the advantage of our re-basin method by proposing a new cost function that allows performing incremental learning by exploiting the linear mode connectivity property. The benefit of our method is compared against similar approaches from the literature under several conditions for both optimal transport and linear mode connectivity. The effectiveness of our continual learning method based on re-basin is also shown for several common benchmark datasets, providing experimental results that are competitive with the state-of-art. The source code is provided at https://github.com/fagp/sinkhorn-rebasin.
****
NerVE: Neural Volumetric Edges for Parametric Curve Extraction From Point Cloud
Xiangyu Zhu, Dong Du, Weikai Chen, Zhiyou Zhao, Yinyu Nie, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13601-13610
Extracting parametric edge curves from point clouds is a fundamental problem in 3D vision and geometry processing. Existing approaches mainly rely on keypoint detection, a challenging procedure that tends to generate noisy output, making the subsequent edge extraction error-prone. To address this issue, we propose to directly detect structured edges to circumvent the limitations of the previous point-wise methods. We achieve this goal by presenting NerVE, a novel neural volumetric edge representation that can be easily learned through a volumetric learning framework. NerVE can be seamlessly converted to a versatile piece-wise linear (PWL) curve representation, enabling a unified strategy for learning all types of free-form curves. Furthermore, as NerVE encodes rich structural information, we show that edge extraction based on NerVE can be reduced to a simple graph search problem. After converting NerVE to the PWL representation, parametric curves can be obtained via off-the-shelf spline fitting algorithms. We evaluate our method on the challenging ABC dataset. We show that a simple network based on NerVE can already outperform the previous state-of-the-art methods by a great margin.
****
ShapeClipper: Scalable 3D Shape Learning From Single-View Images via Geometric and CLIP-Based Consistency
Zixuan Huang, Varun Jampani, Anh Thai, Yuanzhen Li, Stefan Stojanov, James M. Rehg; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12912-12922
We present ShapeClipper, a novel method that reconstructs 3D object shapes from real-world single-view RGB images. Instead of relying on laborious 3D, multi-view or camera pose annotation, ShapeClipper learns shape reconstruction from a set of single-view segmented images. The key idea is to facilitate shape learning via CLIP-based shape consistency, where we encourage objects with similar CLIP en

codings to share similar shapes. We also leverage off-the-shelf normals as an ad ditional geometric constraint so the model can learn better bottom-up reasoning of detailed surface geometry. These two novel consistency constraints, when used to regularize our model, improve its ability to learn both global shape structu re and local geometric details. We evaluate our method over three challenging re al-world datasets, Pix3D, Pascal3D+, and OpenImages, where we achieve superior p erformance over state-of-the-art methods.
****
Supervised Masked Knowledge Distillation for Few-Shot Transformers
Han Lin, Guangxing Han, Jiawei Ma, Shiyuan Huang, Xudong Lin, Shih-Fu Chang; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19649-19659
Vision Transformers (ViTs) emerge to achieve impressive performance on many data -abundant computer vision tasks by capturing long-range dependencies among local features. However, under few-shot learning (FSL) settings on small datasets wit h only a few labeled data, ViT tends to overfit and suffers from severe performa nce degradation due to its absence of CNN-alike inductive bias. Previous works i n FSL avoid such problem either through the help of self-supervised auxiliary lo sses, or through the dextile uses of label information under supervised settings . But the gap between self-supervised and supervised few-shot Transformers is st ill unfilled. Inspired by recent advances in self-supervised knowledge distillat ion and masked image modeling (MIM), we propose a novel Supervised Masked Knowle dge Distillation model (SMKD) for few-shot Transformers which incorporates label information into self-distillation frameworks. Compared with previous self-supe rvised methods, we allow intra-class knowledge distillation on both class and pa tch tokens, and introduce the challenging task of masked patch tokens reconstruc tion across intra-class images. Experimental results on four few-shot classifica tion benchmark datasets show that our method with simple design outperforms prev ious methods by a large margin and achieves a new start-of-the-art. Detailed abl ation studies confirm the effectiveness of each component of our model. Code for this paper is available here: https://github.com/HL-hanlin/SMKD.
****
RIDCP: Revitalizing Real Image Dehazing via High-Quality Codebook Priors
Rui-Qi Wu, Zheng-Peng Duan, Chun-Le Guo, Zhi Chai, Chongyi Li; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 22282-22291
Existing dehazing approaches struggle to process real-world hazy images owing to the lack of paired real data and robust priors. In this work, we present a new paradigm for real image dehazing from the perspectives of synthesizing more real istic hazy data and introducing more robust priors into the network. Specificall y, (1) instead of adopting the de facto physical scattering model, we rethink th e degradation of real hazy images and propose a phenomenological pipeline consid ering diverse degradation types. (2) We propose a Real Image Dehazing network vi a high-quality Codebook Priors (RIDCP). Firstly, a VQGAN is pre-trained on a lar ge-scale high-quality dataset to obtain the discrete codebook, encapsulating hig h-quality priors (HQPs). After replacing the negative effects brought by haze wi th HQPs, the decoder equipped with a novel normalized feature alignment module c an effectively utilize high-quality features and produce clean results. However, although our degradation pipeline drastically mitigates the domain gap between synthetic and real data, it is still intractable to avoid it, which challenges H QPs matching in the wild. Thus, we re-calculate the distance when matching the f eatures to the HQPs by a controllable matching operation, which facilitates find ing better counterparts. We provide a recommendation to control the matching bas ed on an explainable solution. Users can also flexibly adjust the enhancement de gree as per their preference. Extensive experiments verify the effectiveness of our data synthesis pipeline and the superior performance of RIDCP in real image dehazing. Code and data will be released.
****
Exact-NeRF: An Exploration of a Precise Volumetric Parameterization for Neural R adiance Fields

Brian K. S. Isaac-Medina, Chris G. Willcocks, Toby P. Breckon; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 66-75
Neural Radiance Fields (NeRF) have attracted significant attention due to their
ability to synthesize novel scene views with great accuracy. However, inherent t
o their underlying formulation, the sampling of points along a ray with zero wid
th may result in ambiguous representations that lead to further rendering artifa
cts such as aliasing in the final scene. To address this issue, the recent varia
nt mip-NeRF proposes an Integrated Positional Encoding (IPE) based on a conical
view frustum. Although this is expressed with an integral formulation, mip-NeRF
instead approximates this integral as the expected value of a multivariate Gauss
ian distribution. This approximation is reliable for short frustums but degrades
 with highly elongated regions, which arises when dealing with distant scene obj
ects under a larger depth of field. In this paper, we explore the use of an exac
t approach for calculating the IPE by using a pyramid-based integral formulation
 instead of an approximated conical-based one. We denote this formulation as Exa
ct-NeRF and contribute the first approach to offer a precise analytical solution
 to the IPE within the NeRF domain. Our exploratory work illustrates that such a
n exact formulation (Exact-NeRF) matches the accuracy of mip-NeRF and furthermor
e provides a natural extension to more challenging scenarios without further mod
ification, such as in the case of unbounded scenes. Our contribution aims to bot
h address the hitherto unexplored issues of frustum approximation in earlier NeR
F work and additionally provide insight into the potential future consideration
of analytical solutions in future NeRF extensions.
****
Backdoor Attacks Against Deep Image Compression via Adaptive Frequency Trigger
Yi Yu, Yufei Wang, Wenhan Yang, Shijian Lu, Yap-Peng Tan, Alex C. Kot; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2023, pp. 12250-12259
Recent deep-learning-based compression methods have achieved superior performanc
e compared with traditional approaches. However, deep learning models have prove
n to be vulnerable to backdoor attacks, where some specific trigger patterns add
ed to the input can lead to malicious behavior of the models. In this paper, we
present a novel backdoor attack with multiple triggers against learned image com
pression models. Motivated by the widely used discrete cosine transform (DCT) in
 existing compression systems and standards, we propose a frequency-based trigge
r injection model that adds triggers in the DCT domain. In particular, we design
 several attack objectives for various attacking scenarios, including: 1) attack
ing compression quality in terms of bit-rate and reconstruction quality; 2) atta
cking task-driven measures, such as down-stream face recognition and semantic se
gmentation. Moreover, a novel simple dynamic loss is designed to balance the inf
luence of different loss terms adaptively, which helps achieve more efficient tr
aining. Extensive experiments show that with our trained trigger injection model
s and simple modification of encoder parameters (of the compression model), the
proposed attack can successfully inject several backdoors with corresponding tri
ggers in a single image compression model.
****
Recurrence Without Recurrence: Stable Video Landmark Detection With Deep Equilib
rium Models
Paul Micaelli, Arash Vahdat, Hongxu Yin, Jan Kautz, Pavlo Molchanov; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 22814-22825
Cascaded computation, whereby predictions are recurrently refined over several s
tages, has been a persistent theme throughout the development of landmark detect
ion models. In this work, we show that the recently proposed Deep Equilibrium Mo
del (DEQ) can be naturally adapted to this form of computation. Our Landmark DEQ
 (LDEQ) achieves state-of-the-art performance on the challenging WFLW facial lan
dmark dataset, reaching 3.92 NME with fewer parameters and a training memory cos
t of O(1) in the number of recurrent modules. Furthermore, we show that DEQs are
 particularly suited for landmark detection in videos. In this setting, it is ty

pical to train on still images due to the lack of labelled videos. This can lead to a "flickering" effect at inference time on video, whereby a model can rapidly oscillate between different plausible solutions across consecutive frames. By rephrasing DEQs as a constrained optimization, we emulate recurrence at inference time, despite not having access to temporal data at training time. This Recurrence without Recurrence (RwR) paradigm helps in reducing landmark flicker, which we demonstrate by introducing a new metric, normalized mean flicker (NMF), and contributing a new facial landmark video dataset (WFLW-V) targeting landmark uncertainty. On the WFLW-V hard subset made up of 500 videos, our LDEQ with RwR improves the NME and NMF by 10 and 13% respectively, compared to the strongest previously published model using a hand-tuned conventional filter.
****
Generalized Relation Modeling for Transformer Tracking
Shenyuan Gao, Chunluan Zhou, Jun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18686-18695
Compared with previous two-stream trackers, the recent one-stream tracking pipeline, which allows earlier interaction between the template and search region, has achieved a remarkable performance gain. However, existing one-stream trackers always let the template interact with all parts inside the search region throughout all the encoder layers. This could potentially lead to target-background confusion when the extracted feature representations are not sufficiently discriminative. To alleviate this issue, we propose a generalized relation modeling method based on adaptive token division. The proposed method is a generalized formulation of attention-based relation modeling for Transformer tracking, which inherits the merits of both previous two-stream and one-stream pipelines whilst enabling more flexible relation modeling by selecting appropriate search tokens to interact with template tokens. An attention masking strategy and the Gumbel-Softmax technique are introduced to facilitate the parallel computation and end-to-end learning of the token division module. Extensive experiments show that our method is superior to the two-stream and one-stream pipelines and achieves state-of-the-art performance on six challenging benchmarks with a real-time running speed.
****
Non-Line-of-Sight Imaging With Signal Superresolution Network
Jianyu Wang, Xintong Liu, Leping Xiao, Zuoqiang Shi, Lingyun Qiu, Xing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17420-17429
Non-line-of-sight (NLOS) imaging aims at reconstructing the location, shape, albedo, and surface normal of the hidden object around the corner with measured transient data. Due to its strong potential in various fields, it has drawn much attention in recent years. However, long exposure time is not always available for applications such as auto-driving, which hinders the practical use of NLOS imaging. Although scanning fewer points can reduce the total measurement time, it also brings the problem of imaging quality degradation. This paper proposes a general learning-based pipeline for increasing imaging quality with only a few scanning points. We tailor a neural network to learn the operator that recovers a high spatial resolution signal. Experiments on synthetic and measured data indicate that the proposed method provides faithful reconstructions of the hidden scene under both confocal and non-confocal settings. Compared with original measurements, the acquisition of our approach is 16 times faster while maintaining similar reconstruction quality. Besides, the proposed pipeline can be applied directly to existing optical systems and imaging algorithms as a plug-in-and-play module. We believe the proposed pipeline is powerful in increasing the frame rate in NLOS video imaging.
****
WildLight: In-the-Wild Inverse Rendering With a Flashlight
Ziang Cheng, Junxuan Li, Hongdong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4305-4314
This paper proposes a practical photometric solution for the challenging problem of in-the-wild inverse rendering under unknown ambient lighting. Our system recovers scene geometry and reflectance using only multi-view images captured by a

smartphone. The key idea is to exploit smartphone's built-in flashlight as a min imally controlled light source, and decompose image intensities into two photome tric components -- a static appearance corresponds to ambient flux, plus a dynam ic reflection induced by the moving flashlight. Our method does not require flas h/non-flash images to be captured in pairs. Building on the success of neural li ght fields, we use an off-the-shelf method to capture the ambient reflections, w hile the flashlight component enables physically accurate photometric constraint s to decouple reflectance and illumination. Compared to existing inverse renderi ng methods, our setup is applicable to non-darkroom environments yet sidesteps t he inherent difficulties of explicit solving ambient reflections. We demonstrate by extensive experiments that our method is easy to implement, casual to set up , and consistently outperforms existing in-the-wild inverse rendering techniques . Finally, our neural reconstruction can be easily exported to PBR textured tria ngle mesh ready for industrial renderers. Our source code and data are released to https://github.com/za-cheng/WildLight
****
A Probabilistic Attention Model With Occlusion-Aware Texture Regression for 3D H and Reconstruction From a Single RGB Image
Zheheng Jiang, Hossein Rahmani, Sue Black, Bryan M. Williams; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp . 758-767
Recently, deep learning based approaches have shown promising results in 3D hand reconstruction from a single RGB image. These approaches can be roughly divided into model-based approaches, which are heavily dependent on the model's paramet er space, and model-free approaches, which require large numbers of 3D ground tr uths to reduce depth ambiguity and struggle in weakly-supervised scenarios. To o vercome these issues, we propose a novel probabilistic model to achieve the robu stness of model-based approaches and reduced dependence on the model's parameter space of model-free approaches. The proposed probabilistic model incorporates a model-based network as a prior-net to estimate the prior probability distributi on of joints and vertices. An Attention-based Mesh Vertices Uncertainty Regressi on (AMVUR) model is proposed to capture dependencies among vertices and the corr elation between joints and mesh vertices to improve their feature representation . We further propose a learning based occlusion-aware Hand Texture Regression mo del to achieve high-fidelity texture reconstruction. We demonstrate the flexibil ity of the proposed probabilistic model to be trained in both supervised and wea kly-supervised scenarios. The experimental results demonstrate our probabilistic model's state-of-the-art accuracy in 3D hand and texture reconstruction from a single image in both training schemes, including in the presence of severe occlu sions.
****
MixNeRF: Modeling a Ray With Mixture Density for Novel View Synthesis From Spars e Inputs
Seunghyeon Seo, Donghoon Han, Yeonjin Chang, Nojun Kwak; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 206 59-20668
Neural Radiance Field (NeRF) has broken new ground in the novel view synthesis d ue to its simple concept and state-of-the-art quality. However, it suffers from severe performance degradation unless trained with a dense set of images with di fferent camera poses, which hinders its practical applications. Although previou s methods addressing this problem achieved promising results, they relied heavil y on the additional training resources, which goes against the philosophy of spa rse-input novel-view synthesis pursuing the training efficiency. In this work, w e propose MixNeRF, an effective training strategy for novel view synthesis from sparse inputs by modeling a ray with a mixture density model. Our MixNeRF estima tes the joint distribution of RGB colors along the ray samples by modeling it wi th mixture of distributions. We also propose a new task of ray depth estimation as a useful training objective, which is highly correlated with 3D scene geometr y. Moreover, we remodel the colors with regenerated blending weights based on th e estimated ray depth and further improves the robustness for colors and viewpoi

nts. Our MixNeRF outperforms other state-of-the-art methods in various standard benchmarks with superior efficiency of training and inference.
****

A New Path: Scaling Vision-and-Language Navigation With Synthetic Instructions and Imitation Learning
Aishwarya Kamath, Peter Anderson, Su Wang, Jing Yu Koh, Alexander Ku, Austin Waters, Yinfei Yang, Jason Baldridge, Zarana Parekh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10813-10823
Recent studies in Vision-and-Language Navigation (VLN) train RL agents to execute natural-language navigation instructions in photorealistic environments, as a step towards robots that can follow human instructions. However, given the scarcity of human instruction data and limited diversity in the training environments, these agents still struggle with complex language grounding and spatial language understanding. Pre-training on large text and image-text datasets from the web has been extensively explored but the improvements are limited. We investigate large-scale augmentation with synthetic instructions. We take 500+ indoor environments captured in densely-sampled 360 degree panoramas, construct navigation trajectories through these panoramas, and generate a visually-grounded instruction for each trajectory using Marky, a high-quality multilingual navigation instruction generator. We also synthesize image observations from novel viewpoints using an image-to-image GAN. The resulting dataset of 4.2M instruction-trajectory pairs is two orders of magnitude larger than existing human-annotated datasets, and contains a wider variety of environments and viewpoints. To efficiently leverage data at this scale, we train a simple transformer agent with imitation learning. On the challenging RxR dataset, our approach outperforms all existing RL agents, improving the state-of-the-art NDTW from 71.1 to 79.1 in seen environments, and from 64.6 to 66.8 in unseen test environments. Our work points to a new path to improving instruction-following agents, emphasizing large-scale training on near-human quality synthetic instructions.
****

Layout-Based Causal Inference for Object Navigation
Sixian Zhang, Xinhang Song, Weijie Li, Yubing Bai, Xinyao Yu, Shuqiang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10792-10802
Previous works for ObjectNav task attempt to learn the association (e.g. relation graph) between the visual inputs and the goal during training. Such association contains the prior knowledge of navigating in training environments, which is denoted as the experience. The experience performs a positive effect on helping the agent infer the likely location of the goal when the layout gap between the unseen environments of the test and the prior knowledge obtained in training is minor. However, when the layout gap is significant, the experience exerts a negative effect on navigation. Motivated by keeping the positive effect and removing the negative effect of the experience, we propose the layout-based soft Total Direct Effect (L-sTDE) framework based on the causal inference to adjust the prediction of the navigation policy. In particular, we propose to calculate the layout gap which is defined as the KL divergence between the posterior and the prior distribution of the object layout. Then the sTDE is proposed to appropriately control the effect of the experience based on the layout gap. Experimental results on AI2THOR, RoboTHOR, and Habitat demonstrate the effectiveness of our method.
****

Pose-Disentangled Contrastive Learning for Self-Supervised Facial Representation
Yuanyuan Liu, Wenbin Wang, Yibing Zhan, Shaoze Feng, Kejun Liu, Zhe Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9717-9728
Self-supervised facial representation has recently attracted increasing attention due to its ability to perform face understanding without relying on large-scale annotated datasets heavily. However, analytically, current contrastive-based self-supervised learning (SSL) still performs unsatisfactorily for learning facial representation. More specifically, existing contrastive learning (CL) tends to

learn pose-invariant features that cannot depict the pose details of faces, com promising the learning performance. To conquer the above limitation of CL, we pr opose a novel Pose-disentangled Contrastive Learning (PCL) method for general se lf-supervised facial representation. Our PCL first devises a pose-disentangled d ecoder (PDD) with a delicately designed orthogonalizing regulation, which disent angles the pose-related features from the face-aware features; therefore, pose-r elated and other pose-unrelated facial information could be performed in individ ual subnetworks and do not affect each other's training. Furthermore, we introdu ce a pose-related contrastive learning scheme that learns pose-related informati on based on data augmentation of the same image, which would deliver more effect ive face-aware representation for various downstream tasks. We conducted linear evaluation on four challenging downstream facial understanding tasks, i.e., faci al expression recognition, face recognition, AU detection and head pose estimati on.Experimental results demonstrate that PCL significantly outperforms cutting-e dge SSL methods. Our Code is available at https://github.com/DreamMr/PCL.
****
Cross-Domain 3D Hand Pose Estimation With Dual Modalities
Qiuxia Lin, Linlin Yang, Angela Yao; Proceedings of the IEEE/CVF Conference on C omputer Vision and Pattern Recognition (CVPR), 2023, pp. 17184-17193
Recent advances in hand pose estimation have shed light on utilizing synthetic d ata to train neural networks, which however inevitably hinders generalization to real-world data due to domain gaps. To solve this problem, we present a framewo rk for cross-domain semi-supervised hand pose estimation and target the challeng ing scenario of learning models from labelled multi-modal synthetic data and unl abelled real-world data. To that end, we propose a dual-modality network that ex ploits synthetic RGB and synthetic depth images. For pre-training, our network u ses multi-modal contrastive learning and attention-fused supervision to learn ef fective representations of the RGB images. We then integrate a novel self-distil lation technique during fine-tuning to reduce pseudo-label noise. Experiments sh ow that the proposed method significantly improves 3D hand pose estimation and 2 D keypoint detection on benchmarks.
****
Attribute-Preserving Face Dataset Anonymization via Latent Code Optimization
Simone Barattin, Christos Tzelepis, Ioannis Patras, Nicu Sebe; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 8001-8010
This work addresses the problem of anonymizing the identity of faces in a datase t of images, such that the privacy of those depicted is not violated, while at t he same time the dataset is useful for downstream task such as for training mach ine learning models. To the best of our knowledge, we are the first to explicitl y address this issue and deal with two major drawbacks of the existing state-of- the-art approaches, namely that they (i) require the costly training of addition al, purpose-trained neural networks, and/or (ii) fail to retain the facial attri butes of the original images in the anonymized counterparts, the preservation of which is of paramount importance for their use in downstream tasks. We accordin gly present a task-agnostic anonymization procedure that directly optimises the images' latent representation in the latent space of a pre-trained GAN. By optim izing the latent codes directly, we ensure both that the identity is of a desire d distance away from the original (with an identity obfuscation loss), whilst pr eserving the facial attributes (using a novel feature-matching loss in FaRL's de ep feature space). We demonstrate through a series of both qualitative and quant itative experiments that our method is capable of anonymizing the identity of th e images whilst--crucially--better-preserving the facial attributes. We make the code and the pre-trained models publicly available at: https://github.com/chi0t zp/FALCO.
****
Inverse Rendering of Translucent Objects Using Physical and Neural Renderers
Chenhao Li, Trung Thanh Ngo, Hajime Nagahara; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12510-12520
In this work, we propose an inverse rendering model that estimates 3D shape, spa

tially-varying reflectance, homogeneous subsurface scattering parameters, and an environment illumination jointly from only a pair of captured images of a translucent object. In order to solve the ambiguity problem of inverse rendering, we use a physically-based renderer and a neural renderer for scene reconstruction and material editing. Because two renderers are differentiable, we can compute a reconstruction loss to assist parameter estimation. To enhance the supervision of the proposed neural renderer, we also propose an augmented loss. In addition, we use a flash and no-flash image pair as the input. To supervise the training, we constructed a large-scale synthetic dataset of translucent objects, which consists of 117K scenes. Qualitative and quantitative results on both synthetic and real-world datasets demonstrated the effectiveness of the proposed model.
****
Towards Building Self-Aware Object Detectors via Reliable Uncertainty Quantification and Calibration
Kemal Oksuz, Tom Joy, Puneet K. Dokania; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9263-9274
The current approach for testing the robustness of object detectors suffers from serious deficiencies such as improper methods of performing out-of-distribution detection and using calibration metrics which do not consider both localisation and classification quality. In this work, we address these issues, and introduce the Self Aware Object Detection (SAOD) task, a unified testing framework which respects and adheres to the challenges that object detectors face in safety-critical environments such as autonomous driving. Specifically, the SAOD task requires an object detector to be: robust to domain shift; obtain reliable uncertainty estimates for the entire scene; and provide calibrated confidence scores for the detections. We extensively use our framework, which introduces novel metrics and large scale test datasets, to test numerous object detectors in two different use-cases, allowing us to highlight critical insights into their robustness performance. Finally, we introduce a simple baseline for the SAOD task, enabling researchers to benchmark future proposed methods and move towards robust object detectors which are fit for purpose. Code is available at: https://github.com/fiveai/saod
****
Ensemble-Based Blackbox Attacks on Dense Prediction
Zikui Cai, Yaoteng Tan, M. Salman Asif; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4045-4055
We propose an approach for adversarial attacks on dense prediction models (such as object detectors and segmentation). It is well known that the attacks generated by a single surrogate model do not transfer to arbitrary (blackbox) victim models. Furthermore, targeted attacks are often more challenging than the untargeted attacks. In this paper, we show that a carefully designed ensemble can create effective attacks for a number of victim models. In particular, we show that normalization of the weights for individual models plays a critical role in the success of the attacks. We then demonstrate that by adjusting the weights of the ensemble according to the victim model can further improve the performance of the attacks. We performed a number of experiments for object detectors and segmentation to highlight the significance of the our proposed methods. Our proposed ensemble-based method outperforms existing blackbox attack methods for object detection and segmentation. Finally we show that our proposed method can also generate a single perturbation that can fool multiple blackbox detection and segmentation models simultaneously.
****
Improving Fairness in Facial Albedo Estimation via Visual-Textual Cues
Xingyu Ren, Jiankang Deng, Chao Ma, Yichao Yan, Xiaokang Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4511-4520
Recent 3D face reconstruction methods have made significant advances in geometry prediction, yet further cosmetic improvements are limited by lagged albedo because inferring albedo from appearance is an ill-posed problem. Although some existing methods consider prior knowledge from illumination to improve albedo estima

tion, they still produce a light-skin bias due to racially biased albedo models
and limited light constraints. In this paper, we reconsider the relationship bet
ween albedo and face attributes and propose an ID2Albedo to directly estimate al
bedo without constraining illumination. Our key insight is that intrinsic semant
ic attributes such as race, skin color, and age can constrain the albedo map. We
 first introduce visual-textual cues and design a semantic loss to supervise fac
ial albedo estimation. Specifically, we pre-define text labels such as race, ski
n color, age, and wrinkles. Then, we employ the text-image model (CLIP) to compu
te the similarity between the text and the input image, and assign a pseudo-labe
l to each facial image. We constrain generated albedos in the training phase to
have the same attributes as the inputs. In addition, we train a high-quality, un
biased facial albedo generator and utilize the semantic loss to learn the mappin
g from illumination-robust identity features to the albedo latent codes. Finally
, our ID2Albedo is trained in a self-supervised way and outperforms state-of-the
-art albedo estimation methods in terms of accuracy and fidelity. It is worth me
ntioning that our approach has excellent generalizability and fairness, especial
ly on in-the-wild data.
****
Source-Free Video Domain Adaptation With Spatial-Temporal-Historical Consistency
 Learning
Kai Li, Deep Patel, Erik Kruus, Martin Renqiang Min; Proceedings of the IEEE/CVF
 Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14643-1
4652
Source-free domain adaptation (SFDA) is an emerging research topic that studies
how to adapt a pretrained source model using unlabeled target data. It is derive
d from unsupervised domain adaptation but has the advantage of not requiring lab
eled source data to learn adaptive models. This makes it particularly useful in
real-world applications where access to source data is restricted. While there h
as been some SFDA work for images, little attention has been paid to videos. Nai
vely extending image-based methods to videos without considering the unique prop
erties of videos often leads to unsatisfactory results. In this paper, we propos
e a simple and highly flexible method for Source-Free Video Domain Adaptation (S
FVDA), which extensively exploits consistency learning for videos from spatial,
temporal, and historical perspectives. Our method is based on the assumption tha
t videos of the same action category are drawn from the same low-dimensional spa
ce, regardless of the spatio-temporal variations in the high-dimensional space t
hat cause domain shifts. To overcome domain shifts, we simulate spatio-temporal
variations by applying spatial and temporal augmentations on target videos, and
encourage the model to make consistent predictions from a video and its augmente
d versions. Due to the simple design, our method can be applied to various SFVDA
 settings, and experiments show that our method achieves state-of-the-art perfor
mance for all the settings.
****
SmartAssign: Learning a Smart Knowledge Assignment Strategy for Deraining and De
snowing
Yinglong Wang, Chao Ma, Jianzhuang Liu; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3677-3686
Existing methods mainly handle single weather types. However, the connections of
 different weather conditions at deep representation level are usually ignored.
These connections, if used properly, can generate complementary representations
for each other to make up insufficient training data, obtaining positive perform
ance gains and better generalization. In this paper, we focus on the very correl
ated rain and snow to explore their connections at deep representation level. Be
cause sub-optimal connections may cause negative effect, another issue is that i
f rain and snow are handled in a multi-task learning way, how to find an optimal
 connection strategy to simultaneously improve deraining and desnowing performan
ce. To build desired connection, we propose a smart knowledge assignment strateg
y, called SmartAssign, to optimally assign the knowledge learned from both tasks
 to a specific one. In order to further enhance the accuracy of knowledge assign
ment, we propose a novel knowledge contrast mechanism, so that the knowledge ass

igned to different tasks preserves better uniqueness. The inherited inductive bi ases usually limit the modelling ability of CNNs, we introduce a novel transform er block to constitute the backbone of our network to effectively combine long-r ange context dependency and local image details. Extensive experiments on seven benchmark datasets verify that proposed SmartAssign explores effective connectio n between rain and snow, and improves the performances of both deraining and des nowing apparently. The implementation code will be available at https://gitee.co m/mindspore/models/tree/master/research/cv/SmartAssign.
****
Delving Into Discrete Normalizing Flows on SO(3) Manifold for Probabilistic Rota tion Modeling
Yulin Liu, Haoran Liu, Yingda Yin, Yang Wang, Baoquan Chen, He Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 023, pp. 21264-21273
Normalizing flows (NFs) provide a powerful tool to construct an expressive distr ibution by a sequence of trackable transformations of a base distribution and fo rm a probabilistic model of underlying data.Rotation, as an important quantity i n computer vision, graphics, and robotics, can exhibit many ambiguities when occ lusion and symmetry occur and thus demands such probabilistic models. Though muc h progress has been made for NFs in Euclidean space, there are no effective norm alizing flows without discontinuity or many-to-one mapping tailored for SO(3) ma nifold. Given the unique non-Euclidean properties of the rotation manifold, adap ting the existing NFs to SO(3) manifold is non-trivial. In this paper, we propos e a novel normalizing flow on SO(3) by combining a Mobius transformation-based c oupling layer and a quaternion affine transformation. With our proposed rotation normalizing flows, one can not only effectively express arbitrary distributions on SO(3), but also conditionally build the target distribution given input obse rvations. Extensive experiments show that our rotation normalizing flows signifi cantly outperform the baselines on both unconditional and conditional tasks.
****
SfM-TTR: Using Structure From Motion for Test-Time Refinement of Single-View Dep th Networks
Sergio Izquierdo, Javier Civera; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2023, pp. 21466-21476
Estimating a dense depth map from a single view is geometrically ill-posed, and state-of-the-art methods rely on learning depth's relation with visual appearanc e using deep neural networks. On the other hand, Structure from Motion (SfM) lev erages multi-view constraints to produce very accurate but sparse maps, as match ing across images is typically limited by locally discriminative texture. In thi s work, we combine the strengths of both approaches by proposing a novel test-ti me refinement (TTR) method, denoted as SfM-TTR, that boosts the performance of s ingle-view depth networks at test time using SfM multi-view cues. Specifically, and differently from the state of the art, we use sparse SfM point clouds as tes t-time self-supervisory signal, fine-tuning the network encoder to learn a bette r representation of the test scene. Our results show how the addition of SfM-TTR to several state-of-the-art self-supervised and supervised networks improves si gnificantly their performance, outperforming previous TTR baselines mainly based on photometric multi-view consistency. The code is available at https://github. com/serizba/SfM-TTR.
****
Fusing Pre-Trained Language Models With Multimodal Prompts Through Reinforcement Learning
Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, R owan Zellers, Prithviraj Ammanabrolu, Ronan Le Bras, Gunhee Kim, Yejin Choi; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10845-10856
Language models are capable of commonsense reasoning: while domain-specific mode ls can learn from explicit knowledge (e.g. commonsense graphs [6], ethical norms [25]), and larger models like GPT-3 manifest broad commonsense reasoning capaci ty. Can their knowledge be extended to multimodal inputs such as images and audi

o without paired domain data? In this work, we propose ESPER (Extending Sensory PErception with Reinforcement learning) which enables text-only pretrained models to address multimodal tasks such as visual commonsense reasoning. Our key novelty is to use reinforcement learning to align multimodal inputs to language model generations without direct supervision: for example, our reward optimization relies only on cosine similarity derived from CLIP and requires no additional paired (image, text) data. Experiments demonstrate that ESPER outperforms baselines and prior work on a variety of multimodal text generation tasks ranging from captioning to commonsense reasoning; these include a new benchmark we collect and release, the ESP dataset, which tasks models with generating the text of several different domains for each image. Our code and data are publicly released at https://github.com/JiwanChung/esper.
****
MELTR: Meta Loss Transformer for Learning To Fine-Tune Video Foundation Models
Dohwan Ko, Joonmyung Choi, Hyeong Kyu Choi, Kyoung-Woon On, Byungseok Roh, Hyunwoo J. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20105-20115
Foundation models have shown outstanding performance and generalization capabilities across domains. Since most studies on foundation models mainly focus on the pretraining phase, a naive strategy to minimize a single task-specific loss is adopted for fine-tuning. However, such fine-tuning methods do not fully leverage other losses that are potentially beneficial for the target task. Therefore, we propose MEta Loss TRansformer (MELTR), a plug-in module that automatically and non-linearly combines various loss functions to aid learning the target task via auxiliary learning. We formulate the auxiliary learning as a bi-level optimization problem and present an efficient optimization algorithm based on Approximate Implicit Differentiation (AID). For evaluation, we apply our framework to various video foundation models (UniVL, Violet and All-in-one), and show significant performance gain on all four downstream tasks: text-to-video retrieval, video question answering, video captioning, and multi-modal sentiment analysis. Our qualitative analyses demonstrate that MELTR adequately 'transforms' individual loss functions and 'melts' them into an effective unified loss. Code is available at https://github.com/mlvlab/MELTR.
****
Dense Network Expansion for Class Incremental Learning
Zhiyuan Hu, Yunsheng Li, Jiancheng Lyu, Dashan Gao, Nuno Vasconcelos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11858-11867
The problem of class incremental learning (CIL) is considered. State-of-the-art approaches use a dynamic architecture based on network expansion (NE), in which a task expert is added per task. While effective from a computational standpoint, these methods lead to models that grow quickly with the number of tasks. A new NE method, dense network expansion (DNE), is proposed to achieve a better trade-off between accuracy and model complexity. This is accomplished by the introduction of dense connections between the intermediate layers of the task expert networks, that enable the transfer of knowledge from old to new tasks via feature sharing and reusing. This sharing is implemented with a cross-task attention mechanism, based on a new task attention block (TAB), that fuses information across tasks. Unlike traditional attention mechanisms, TAB operates at the level of the feature mixing and is decoupled with spatial attentions. This is shown more effective than a joint spatial-and-task attention for CIL. The proposed DNE approach can strictly maintain the feature space of old classes while growing the network and feature scale at a much slower rate than previous methods. In result, it outperforms the previous SOTA methods by a margin of 4% in terms of accuracy, with similar or even smaller model scale.
****
Meta-Personalizing Vision-Language Models To Find Named Instances in Video
Chun-Hsiao Yeh, Bryan Russell, Josef Sivic, Fabian Caba Heilbron, Simon Jenni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19123-19132

Large-scale vision-language models (VLM) have shown impressive results for language-guided search applications. While these models allow category-level queries, they currently struggle with personalized searches for moments in a video where a specific object instance such as "My dog Biscuit" appears. We present the following three contributions to address this problem. First, we describe a method to meta-personalize a pre-trained VLM, i.e., learning how to learn to personalize a VLM at test time to search in video. Our method extends the VLM's token vocabulary by learning novel word embeddings specific to each instance. To capture only instance-specific features, we represent each instance embedding as a combination of shared and learned global category features. Second, we propose to learn such personalization without explicit human supervision. Our approach automatically identifies moments of named visual instances in video using transcripts and vision-language similarity in the VLM's embedding space. Finally, we introduce This-Is-My, a personal video instance retrieval benchmark. We evaluate our approach on This-Is-My and DeepFashion2 and show that we obtain a 15% relative improvement over the state of the art on the latter dataset.
****

Regularize Implicit Neural Representation by Itself
Zhemin Li, Hongxia Wang, Deyu Meng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10280-10288
This paper proposes a regularizer called Implicit Neural Representation Regularizer (INRR) to improve the generalization ability of the Implicit Neural Representation (INR). The INR is a fully connected network that can represent signals with details not restricted by grid resolution. However, its generalization ability could be improved, especially with non-uniformly sampled data. The proposed INRR is based on learned Dirichlet Energy (DE) that measures similarities between rows/columns of the matrix. The smoothness of the Laplacian matrix is further integrated by parameterizing DE with a tiny INR. INRR improves the generalization of INR in signal representation by perfectly integrating the signal's self-similarity with the smoothness of the Laplacian matrix. Through well-designed numerical experiments, the paper also reveals a series of properties derived from INRR, including momentum methods like convergence trajectory and multi-scale similarity. Moreover, the proposed method could improve the performance of other signal representation methods.
****

Egocentric Audio-Visual Object Localization
Chao Huang, Yapeng Tian, Anurag Kumar, Chenliang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22910-22921
Humans naturally perceive surrounding scenes by unifying sound and sight in a first-person view. Likewise, machines are advanced to approach human intelligence by learning with multisensory inputs from an egocentric perspective. In this paper, we explore the challenging egocentric audio-visual object localization task and observe that 1) egomotion commonly exists in first-person recordings, even within a short duration; 2) The out-of-view sound components can be created while wearers shift their attention. To address the first problem, we propose a geometry-aware temporal aggregation module to handle the egomotion explicitly. The effect of egomotion is mitigated by estimating the temporal geometry transformation and exploiting it to update visual representations. Moreover, we propose a cascaded feature enhancement module to tackle the second issue. It improves cross-modal localization robustness by disentangling visually-indicated audio representation. During training, we take advantage of the naturally available audio-visual temporal synchronization as the "free" self-supervision to avoid costly labeling. We also annotate and create the Epic Sounding Object dataset for evaluation purposes. Extensive experiments show that our method achieves state-of-the-art localization performance in egocentric videos and can be generalized to diverse audio-visual scenes.
****

DropKey for Vision Transformer
Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, Lu

oqi Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22700-22709
In this paper, we focus on analyzing and improving the dropout technique for self-attention layers of Vision Transformer, which is important while surprisingly ignored by prior works. In particular, we conduct researches on three core questions: First, what to drop in self-attention layers? Different from dropping attention weights in literature, we propose to move dropout operations forward ahead of attention matrix calculation and set the Key as the dropout unit, yielding a novel dropout-before-softmax scheme. We theoretically verify that this scheme helps keep both regularization and probability features of attention weights, alleviating the overfittings problem to specific patterns and enhancing the model to globally capture vital information; Second, how to schedule the drop ratio in consecutive layers? In contrast to exploit a constant drop ratio for all layers, we present a new decreasing schedule that gradually decreases the drop ratio along the stack of self-attention layers. We experimentally validate the proposed schedule can avoid overfittings in low-level features and missing in high-level semantics, thus improving the robustness and stableness of model training; Third, whether need to perform structured dropout operation as CNN? We attempt patch-based block-version of dropout operation and find that this useful trick for CNN is not essential for ViT. Given exploration on the above three questions, we present the novel DropKey method that regards Key as the drop unit and exploits decreasing schedule for drop ratio, improving ViTs in a general way. Comprehensive experiments demonstrate the effectiveness of DropKey for various ViT architectures, e.g. T2T, VOLO, CeiT and DeiT, as well as for various vision tasks, e.g., image classification, object detection, human-object interaction detection and human body shape recovery.
****
sRGB Real Noise Synthesizing With Neighboring Correlation-Aware Noise Model
Zixuan Fu, Lanqing Guo, Bihan Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1683-1691
Modeling and synthesizing real noise in the standard RGB (sRGB) domain is challenging due to the complicated noise distribution. While most of the deep noise generators proposed to synthesize sRGB real noise using an end-to-end trained model, the lack of explicit noise modeling degrades the quality of their synthesized noise. In this work, we propose to model the real noise as not only dependent on the underlying clean image pixel intensity, but also highly correlated to its neighboring noise realization within the local region. Correspondingly, we propose a novel noise synthesizing framework by explicitly learning its neighboring correlation on top of the signal dependency. With the proposed noise model, our framework greatly bridges the distribution gap between synthetic noise and real noise. We show that our generated "real" sRGB noisy images can be used for training supervised deep denoisers, thus to improve their real denoising results with a large margin, comparing to the popular classic denoisers or the deep denoisers that are trained on other sRGB noise generators. The code will be available at https://github.com/xuan611/sRGB-Real-Noise-Synthesizing.
****
Meta Architecture for Point Cloud Analysis
Haojia Lin, Xiawu Zheng, Lijiang Li, Fei Chao, Shanshan Wang, Yan Wang, Yonghong Tian, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17682-17691
Recent advances in 3D point cloud analysis bring a diverse set of network architectures to the field. However, the lack of a unified framework to interpret those networks makes any systematic comparison, contrast, or analysis challenging, and practically limits healthy development of the field. In this paper, we take the initiative to explore and propose a unified framework called PointMeta, to which the popular 3D point cloud analysis approaches could fit. This brings three benefits. First, it allows us to compare different approaches in a fair manner, and use quick experiments to verify any empirical observations or assumptions summarized from the comparison. Second, the big picture brought by PointMeta enables us to think across different components, and revisit common beliefs and key d

esign decisions made by the popular approaches. Third, based on the learnings fr
om the previous two analyses, by doing simple tweaks on the existing approaches,
 we are able to derive a basic building block, termed PointMetaBase. It shows ve
ry strong performance in efficiency and effectiveness through extensive experime
nts on challenging benchmarks, and thus verifies the necessity and benefits of h
igh-level interpretation, contrast, and comparison like PointMeta. In particular
, PointMetaBase surpasses the previous state-of-the-art method by 0.7%/1.4/%2.1%
 mIoU with only 2%/11%/13% of the computation cost on the S3DIS datasets. Codes
are available in the supplementary materials.
****
Ambiguous Medical Image Segmentation Using Diffusion Models
Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2023, pp. 11536-11546
Collective insights from a group of experts have always proven to outperform an
individual's best diagnostic for clinical tasks. For the task of medical image s
egmentation, existing research on AI-based alternatives focuses more on developi
ng models that can imitate the best individual rather than harnessing the power
of expert groups. In this paper, we introduce a single diffusion model-based app
roach that produces multiple plausible outputs by learning a distribution over g
roup insights. Our proposed model generates a distribution of segmentation masks
 by leveraging the inherent stochastic sampling process of diffusion using only
minimal additional learning. We demonstrate on three different medical image mod
alities- CT, ultrasound, and MRI that our model is capable of producing several
possible variants while capturing the frequencies of their occurrences. Comprehe
nsive results show that our proposed approach outperforms existing state-of-the-
art ambiguous segmentation networks in terms of accuracy while preserving natura
lly occurring variation. We also propose a new metric to evaluate the diversity
as well as the accuracy of segmentation predictions that aligns with the interes
t of clinical practice of collective insights. Implementation code will be relea
sed publicly after the review process.
****
CIRCLE: Capture in Rich Contextual Environments
João Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak
 Gopinath, Alexander William Clegg, Karen Liu; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21211-21221
Synthesizing 3D human motion in a contextual, ecological environment is importan
t for simulating realistic activities people perform in the real world. However,
 conventional optics-based motion capture systems are not suited for simultaneou
sly capturing human movements and complex scenes. The lack of rich contextual 3D
 human motion datasets presents a roadblock to creating high-quality generative
human motion models. We propose a novel motion acquisition system in which the a
ctor perceives and operates in a highly contextual virtual world while being mot
ion captured in the real world. Our system enables rapid collection of high-qual
ity human motion in highly diverse scenes, without the concern of occlusion or t
he need for physical scene construction in the real world. We present CIRCLE, a
dataset containing 10 hours of full-body reaching motion from 5 subjects across
nine scenes, paired with ego-centric information of the environment represented
in various forms, such as RGBD videos. We use this dataset to train a model that
 generates human motion conditioned on scene information. Leveraging our dataset
, the model learns to use ego-centric scene information to achieve nontrivial re
aching tasks in the context of complex 3D scenes. To download the data please vi
sit our website (https://stanford-tml.github.io/circle_dataset/).
****
Revisiting Weak-to-Strong Consistency in Semi-Supervised Semantic Segmentation
Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, Yinghuan Shi; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7
236-7246
In this work, we revisit the weak-to-strong consistency framework, popularized b
y FixMatch from semi-supervised classification, where the prediction of a weakly

perturbed image serves as supervision for its strongly perturbed version. Intriguingly, we observe that such a simple pipeline already achieves competitive results against recent advanced works, when transferred to our segmentation scenario. Its success heavily relies on the manual design of strong data augmentations, however, which may be limited and inadequate to explore a broader perturbation space. Motivated by this, we propose an auxiliary feature perturbation stream as a supplement, leading to an expanded perturbation space. On the other, to sufficiently probe original image-level augmentations, we present a dual-stream perturbation technique, enabling two strong views to be simultaneously guided by a common weak view. Consequently, our overall Unified Dual-Stream Perturbations approach (UniMatch) surpasses all existing methods significantly across all evaluation protocols on the Pascal, Cityscapes, and COCO benchmarks. Its superiority is also demonstrated in remote sensing interpretation and medical image analysis. We hope our reproduced FixMatch and our results can inspire more future works. Code and logs are available at https://github.com/LiheYoung/UniMatch.
****
Implicit View-Time Interpolation of Stereo Videos Using Multi-Plane Disparities and Non-Uniform Coordinates
Avinash Paliwal, Andrii Tsarov, Nima Khademi Kalantari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 888-898
In this paper, we propose an approach for view-time interpolation of stereo videos. Specifically, we build upon X-Fields that approximates an interpolatable mapping between the input coordinates and 2D RGB images using a convolutional decoder. Our main contribution is to analyze and identify the sources of the problems with using X-Fields in our application and propose novel techniques to overcome these challenges. Specifically, we observe that X-Fields struggles to implicitly interpolate the disparities for large baseline cameras. Therefore, we propose multi-plane disparities to reduce the spatial distance of the objects in the stereo views. Moreover, we propose non-uniform time coordinates to handle the non-linear and sudden motion spikes in videos. We additionally introduce several simple, but important, improvements over X-Fields. We demonstrate that our approach is able to produce better results than the state of the art, while running in near real-time rates and having low memory and storage costs.
****
PyPose: A Library for Robot Learning With Physics-Based Optimization
Chen Wang, Dasong Gao, Kuan Xu, Junyi Geng, Yaoyu Hu, Yuheng Qiu, Bowen Li, Fan Yang, Brady Moon, Abhinav Pandey, Aryan, Jiahe Xu, Tianhao Wu, Haonan He, Daning Huang, Zhongqiang Ren, Shibo Zhao, Taimeng Fu, Pranay Reddy, Xiao Lin, Wenshan Wang, Jingnan Shi, Rajat Talak, Kun Cao, Yi Du, Han Wang, Huai Yu, Shanzhao Wang, Siyu Chen, Ananth Kashyap, Rohan Bandaru, Karthik Dantu, Jiajun Wu, Lihua Xie, Luca Carlone, Marco Hutter, Sebastian Scherer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22024-22034
Deep learning has had remarkable success in robotic perception, but its data-centric nature suffers when it comes to generalizing to ever-changing environments. By contrast, physics-based optimization generalizes better, but it does not perform as well in complicated tasks due to the lack of high-level semantic information and reliance on manual parametric tuning. To take advantage of these two complementary worlds, we present PyPose: a robotics-oriented, PyTorch-based library that combines deep perceptual models with physics-based optimization. PyPose's architecture is tidy and well-organized, it has an imperative style interface and is efficient and user-friendly, making it easy to integrate into real-world robotic applications. Besides, it supports parallel computing of any order gradients of Lie groups and Lie algebras and 2nd-order optimizers, such as trust region methods. Experiments show that PyPose achieves more than 10x speedup in computation compared to the state-of-the-art libraries. To boost future research, we provide concrete examples for several fields of robot learning, including SLAM, planning, control, and inertial navigation.
****
Make Landscape Flatter in Differentially Private Federated Learning

Yifan Shi, Yingqi Liu, Kang Wei, Li Shen, Xueqian Wang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24552-24562

To defend the inference attacks and mitigate the sensitive information leakages in Federated Learning (FL), client-level Differentially Private FL (DPFL) is the de-facto standard for privacy protection by clipping local updates and adding random noise. However, existing DPFL methods tend to make a sharper loss landscape and have poorer weight perturbation robustness, resulting in severe performance degradation. To alleviate these issues, we propose a novel DPFL algorithm named DP-FedSAM, which leverages gradient perturbation to mitigate the negative impact of DP. Specifically, DP-FedSAM integrates Sharpness Aware Minimization (SAM) optimizer to generate local flatness models with better stability and weight perturbation robustness, which results in the small norm of local updates and robustness to DP noise, thereby improving the performance. From the theoretical perspective, we analyze in detail how DP-FedSAM mitigates the performance degradation induced by DP. Meanwhile, we give rigorous privacy guarantees with Renyi DP and present the sensitivity analysis of local updates. At last, we empirically confirm that our algorithm achieves state-of-the-art (SOTA) performance compared with existing SOTA baselines in DPFL.

****

BlackVIP: Black-Box Visual Prompting for Robust Transfer Learning

Changdae Oh, Hyeji Hwang, Hee-young Lee, YongTaek Lim, Geunyoung Jung, Jiyoung Jung, Hosik Choi, Kyungwoo Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24224-24235

With the surge of large-scale pre-trained models (PTMs), fine-tuning these models to numerous downstream tasks becomes a crucial problem. Consequently, parameter efficient transfer learning (PETL) of large models has grasped huge attention. While recent PETL methods showcase impressive performance, they rely on optimistic assumptions: 1) the entire parameter set of a PTM is available, and 2) a sufficiently large memory capacity for the fine-tuning is equipped. However, in most real-world applications, PTMs are served as a black-box API or proprietary software without explicit parameter accessibility. Besides, it is hard to meet a large memory requirement for modern PTMs. In this work, we propose black-box visual prompting (BlackVIP), which efficiently adapts the PTMs without knowledge about model architectures and parameters. BlackVIP has two components; 1) Coordinator and 2) simultaneous perturbation stochastic approximation with gradient correction (SPSA-GC). The Coordinator designs input-dependent image-shaped visual prompts, which improves few-shot adaptation and robustness on distribution/location shift. SPSA-GC efficiently estimates the gradient of a target model to update Coordinator. Extensive experiments on 16 datasets demonstrate that BlackVIP enables robust adaptation to diverse domains without accessing PTMs' parameters, with minimal memory requirements. Code: https://github.com/changdaeoh/BlackVIP

****

DeepVecFont-v2: Exploiting Transformers To Synthesize Vector Fonts With Higher Quality

Yuqing Wang, Yizhi Wang, Longhui Yu, Yuesheng Zhu, Zhouhui Lian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18320-18328

Vector font synthesis is a challenging and ongoing problem in the fields of Computer Vision and Computer Graphics. The recently-proposed DeepVecFont achieved state-of-the-art performance by exploiting information of both the image and sequence modalities of vector fonts. However, it has limited capability for handling long sequence data and heavily relies on an image-guided outline refinement post-processing. Thus, vector glyphs synthesized by DeepVecFont still often contain some distortions and artifacts and cannot rival human-designed results. To address the above problems, this paper proposes an enhanced version of DeepVecFont mainly by making the following three novel technical contributions. First, we adopt Transformers instead of RNNs to process sequential data and design a relaxation representation for vector outlines, markedly improving the model's capability and stability of synthesizing long and complex outlines. Second, we propose to s

ample auxiliary points in addition to control points to precisely align the generated and target Bezier curves or lines. Finally, to alleviate error accumulation in the sequential generation process, we develop a context-based self-refinement module based on another Transformer-based decoder to remove artifacts in the initially synthesized glyphs. Both qualitative and quantitative results demonstrate that the proposed method effectively resolves those intrinsic problems of the original DeepVecFont and outperforms existing approaches in generating English and Chinese vector fonts with complicated structures and diverse styles.
****
pCON: Polarimetric Coordinate Networks for Neural Scene Representations
Henry Peters, Yunhao Ba, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16579-16589
Neural scene representations have achieved great success in parameterizing and reconstructing images, but current state of the art models are not optimized with the preservation of physical quantities in mind. While current architectures can reconstruct color images correctly, they create artifacts when trying to fit maps of polar quantities. We propose polarimetric coordinate networks (pCON), a new model architecture for neural scene representations aimed at preserving polarimetric information while accurately parameterizing the scene. Our model removes artifacts created by current coordinate network architectures when reconstructing three polarimetric quantities of interest.
****
Soft-Landing Strategy for Alleviating the Task Discrepancy Problem in Temporal Action Localization Tasks
Hyolim Kang, Hanjung Kim, Joungbin An, Minsu Cho, Seon Joo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6514-6523
Temporal Action Localization (TAL) methods typically operate on top of feature sequences from a frozen snippet encoder that is pretrained with the Trimmed Action Classification (TAC) tasks, resulting in a task discrepancy problem. While existing TAL methods mitigate this issue either by retraining the encoder with a pretext task or by end-to-end finetuning, they commonly require an overload of high memory and computation. In this work, we introduce Soft-Landing (SoLa) strategy, an efficient yet effective framework to bridge the transferability gap between the pretrained encoder and the downstream tasks by incorporating a light-weight neural network, i.e., a SoLa module, on top of the frozen encoder. We also propose an unsupervised training scheme for the SoLa module; it learns with inter-frame Similarity Matching that uses the frame interval as its supervisory signal, eliminating the need for temporal annotations. Experimental evaluation on various benchmarks for downstream TAL tasks shows that our method effectively alleviates the task discrepancy problem with remarkable computational efficiency.
****
Visibility Aware Human-Object Interaction Tracking From Single RGB Camera
Xianghui Xie, Bharat Lal Bhatnagar, Gerard Pons-Moll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4757-4768
Capturing the interactions between humans and their environment in 3D is important for many applications in robotics, graphics, and vision. Recent works to reconstruct the 3D human and object from a single RGB image do not have consistent relative translation across frames because they assume a fixed depth. Moreover, their performance drops significantly when the object is occluded. In this work, we propose a novel method to track the 3D human, object, contacts, and relative translation across frames from a single RGB camera, while being robust to heavy occlusions. Our method is built on two key insights. First, we condition our neural field reconstructions for human and object on per-frame SMPL model estimates obtained by pre-fitting SMPL to a video sequence. This improves neural reconstruction accuracy and produces coherent relative translation across frames. Second, human and object motion from visible frames provides valuable information to infer the occluded object. We propose a novel transformer-based neural network that explicitly uses object visibility and human motion to leverage neighboring fr

ames to make predictions for the occluded frames. Building on these insights, ou
r method is able to track both human and object robustly even under occlusions.
Experiments on two datasets show that our method significantly improves over the
 state-of-the-art methods. Our code and pretrained models are available at: http
s://virtualhumans.mpi-inf.mpg.de/VisTracker.
****
Uncertainty-Aware Vision-Based Metric Cross-View Geolocalization
Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, Rai
ner Stiefelhagen; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2023, pp. 21621-21631
This paper proposes a novel method for vision-based metric cross-view geolocaliz
ation (CVGL) that matches the camera images captured from a ground-based vehicle
 with an aerial image to determine the vehicle's geo-pose. Since aerial images a
re globally available at low cost, they represent a potential compromise between
 two established paradigms of autonomous driving, i.e. using expensive high-defi
nition prior maps or relying entirely on the sensor data captured at runtime. We
 present an end-to-end differentiable model that uses the ground and aerial imag
es to predict a probability distribution over possible vehicle poses. We combine
 multiple vehicle datasets with aerial images from orthophoto providers on which
 we demonstrate the feasibility of our method. Since the ground truth poses are
often inaccurate w.r.t. the aerial images, we implement a pseudo-label approach
to produce more accurate ground truth poses and make them publicly available. Wh
ile previous works require training data from the target region to achieve reaso
nable localization accuracy (i.e. same-area evaluation), our approach overcomes
this limitation and outperforms previous results even in the strictly more chall
enging cross-area case. We improve the previous state-of-the-art by a large marg
in even without ground or aerial data from the test region, which highlights the
 model's potential for global-scale application. We further integrate the uncert
ainty-aware predictions in a tracking framework to determine the vehicle's traje
ctory over time resulting in a mean position error on KITTI-360 of 0.78m.
****
DANI-Net: Uncalibrated Photometric Stereo by Differentiable Shadow Handling, Ani
sotropic Reflectance Modeling, and Neural Inverse Rendering
Zongrui Li, Qian Zheng, Boxin Shi, Gang Pan, Xudong Jiang; Proceedings of the IE
EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8
381-8391
Uncalibrated photometric stereo (UPS) is challenging due to the inherent ambigui
ty brought by the unknown light. Although the ambiguity is alleviated on non-Lam
bertian objects, the problem is still difficult to solve for more general object
s with complex shapes introducing irregular shadows and general materials with c
omplex reflectance like anisotropic reflectance. To exploit cues from shadow and
 reflectance to solve UPS and improve performance on general materials, we propo
se DANI-Net, an inverse rendering framework with differentiable shadow handling
and anisotropic reflectance modeling. Unlike most previous methods that use non-
differentiable shadow maps and assume isotropic material, our network benefits f
rom cues of shadow and anisotropic reflectance through two differentiable paths.
 Experiments on multiple real-world datasets demonstrate our superior and robust
 performance.
****
Towards Better Stability and Adaptability: Improve Online Self-Training for Mode
l Adaptation in Semantic Segmentation
Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, Licheng Jiao; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 11733-11743
Unsupervised domain adaptation (UDA) in semantic segmentation transfers the know
ledge of the source domain to the target one to improve the adaptability of the
segmentation model in the target domain. The need to access labeled source data
makes UDA unable to handle adaptation scenarios involving privacy, property righ
ts protection, and confidentiality. In this paper, we focus on unsupervised mode
l adaptation (UMA), also called source-free domain adaptation, which adapts a so

urce-trained model to the target domain without accessing source data. We find t
hat the online self-training method has the potential to be deployed in UMA, but
 the lack of source domain loss will greatly weaken the stability and adaptabili
ty of the method. We analyze the two possible reasons for the degradation of onl
ine self-training, i.e. inopportune updates of the teacher model and biased know
ledge from source-trained model. Based on this, we propose a dynamic teacher upd
ate mechanism and a training-consistency based resampling strategy to improve th
e stability and adaptability of online self training. On multiple model adaptati
on benchmarks, our method obtains new state-of-the-art performance, which is com
parable or even better than state-of-the-art UDA methods.
****
Continuous Landmark Detection With 3D Queries
Prashanth Chandran, Gaspard Zoss, Paulo Gotardo, Derek Bradley; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 16858-16867
Neural networks for facial landmark detection are notoriously limited to a fixed
 set of landmarks in a dedicated layout, which must be specified at training tim
e. Dedicated datasets must also be hand-annotated with the corresponding landmar
k configuration for training. We propose the first facial landmark detection net
work that can predict continuous, unlimited landmarks, allowing to specify the n
umber and location of the desired landmarks at inference time. Our method combin
es a simple image feature extractor with a queried landmark predictor, and the u
ser can specify any continuous query points relative to a 3D template face mesh
as input. As it is not tied to a fixed set of landmarks, our method is able to l
everage all pre-existing 2D landmark datasets for training, even if they have in
consistent landmark configurations. As a result, we present a very powerful faci
al landmark detector that can be trained once, and can be used readily for numer
ous applications like 3D face reconstruction, arbitrary face segmentation, and i
s even compatible with helmeted mounted cameras, and therefore could vastly simp
lify face tracking workflows for media and entertainment applications.
****
Ranking Regularization for Critical Rare Classes: Minimizing False Positives at
a High True Positive Rate
Kiarash Mohammadi, He Zhao, Mengyao Zhai, Frederick Tung; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15
783-15792
In many real-world settings, the critical class is rare and a missed detection c
arries a disproportionately high cost. For example, tumors are rare and a false
negative diagnosis could have severe consequences on treatment outcomes; fraudul
ent banking transactions are rare and an undetected occurrence could result in s
ignificant losses or legal penalties. In such contexts, systems are often operat
ed at a high true positive rate, which may require tolerating high false positiv
es. In this paper, we present a novel approach to address the challenge of minim
izing false positives for systems that need to operate at a high true positive r
ate. We propose a ranking-based regularization (RankReg) approach that is easy t
o implement, and show empirically that it not only effectively reduces false pos
itives, but also complements conventional imbalanced learning losses. With this
novel technique in hand, we conduct a series of experiments on three broadly exp
lored datasets (CIFAR-10&100 and Melanoma) and show that our approach lifts the
previous state-of-the-art performance by notable margins.
****
Rethinking Gradient Projection Continual Learning: Stability / Plasticity Featur
e Space Decoupling
Zhen Zhao, Zhizhong Zhang, Xin Tan, Jun Liu, Yanyun Qu, Yuan Xie, Lizhuang Ma; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2023, pp. 3718-3727
Continual learning aims to incrementally learn novel classes over time, while no
t forgetting the learned knowledge. Recent studies have found that learning woul
d not forget if the updated gradient is orthogonal to the feature space. However
, previous approaches require the gradient to be fully orthogonal to the whole f

eature space, leading to poor plasticity, as the feasible gradient direction becomes narrow when the tasks continually come, i.e., feature space is unlimitedly expanded. In this paper, we propose a space decoupling (SD) algorithm to decouple the feature space into a pair of complementary subspaces, i.e., the stability space I, and the plasticity space R. I is established by conducting space intersection between the historic and current feature space, and thus I contains more task-shared bases. R is constructed by seeking the orthogonal complementary subspace of I, and thus R mainly contains more task-specific bases. By putting the distinguishing constraints on R and I, our method achieves a better balance between stability and plasticity. Extensive experiments are conducted by applying SD to gradient projection baselines, and show SD is model-agnostic and achieves SOTA results on publicly available datasets.
****
Joint HDR Denoising and Fusion: A Real-World Mobile HDR Image Dataset
Shuaizheng Liu, Xindong Zhang, Lingchen Sun, Zhetong Liang, Hui Zeng, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13966-13975
Mobile phones have become a ubiquitous and indispensable photographing device in our daily life, while the small aperture and sensor size make mobile phones more susceptible to noise and over-saturation, resulting in low dynamic range (LDR) and low image quality. It is thus crucial to develop high dynamic range (HDR) imaging techniques for mobile phones. Unfortunately, the existing HDR image datasets are mostly constructed by DSLR cameras in daytime, limiting their applicability to the study of HDR imaging for mobile phones. In this work, we develop, for the first time to our best knowledge, an HDR image dataset by using mobile phone cameras, namely Mobile-HDR dataset. Specifically, we utilize three mobile phone cameras to collect paired LDR-HDR images in the raw image domain, covering both daytime and nighttime scenes with different noise levels. We then propose a transformer based model with a pyramid cross-attention alignment module to aggregate highly correlated features from different exposure frames to perform joint HDR denoising and fusion. Experiments validate the advantages of our dataset and our method on mobile HDR imaging. Dataset and codes are available at https://github.com/shuaizhengliu/Joint-HDRDN.
****
FlatFormer: Flattened Window Attention for Efficient Point Cloud Transformer
Zhijian Liu, Xinyu Yang, Haotian Tang, Shang Yang, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1200-1211
Transformer, as an alternative to CNN, has been proven effective in many modalities (e.g., texts and images). For 3D point cloud transformers, existing efforts focus primarily on pushing their accuracy to the state-of-the-art level. However, their latency lags behind sparse convolution-based models (3x slower), hindering their usage in resource-constrained, latency-sensitive applications (such as autonomous driving). This inefficiency comes from point clouds' sparse and irregular nature, whereas transformers are designed for dense, regular workloads. This paper presents FlatFormer to close this latency gap by trading spatial proximity for better computational regularity. We first flatten the point cloud with window-based sorting and partition points into groups of equal sizes rather than windows of equal shapes. This effectively avoids expensive structuring and padding overheads. We then apply self-attention within groups to extract local features, alternate sorting axis to gather features from different directions, and shift windows to exchange features across groups. FlatFormer delivers state-of-the-art accuracy on Waymo Open Dataset with 4.6x speedup over (transformer-based) SST and 1.4x speedup over (sparse convolutional) CenterPoint. This is the first point cloud transformer that achieves real-time performance on edge GPUs and is faster than sparse convolutional methods while achieving on-par or even superior accuracy on large-scale benchmarks.
****
Unbiased Scene Graph Generation in Videos
Sayak Nag, Kyle Min, Subarna Tripathi, Amit K. Roy-Chowdhury; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22803-22813
The task of dynamic scene graph generation (SGG) from videos is complicated and challenging due to the inherent dynamics of a scene, temporal fluctuation of model predictions, and the long-tailed distribution of the visual relationships in addition to the already existing challenges in image-based SGG. Existing methods for dynamic SGG have primarily focused on capturing spatio-temporal context using complex architectures without addressing the challenges mentioned above, especially the long-tailed distribution of relationships. This often leads to the generation of biased scene graphs. To address these challenges, we introduce a new framework called TEMPURA: TEmporal consistency and Memory Prototype guided UnceRtainty Attenuation for unbiased dynamic SGG. TEMPURA employs object-level temporal consistencies via transformer-based sequence modeling, learns to synthesize unbiased relationship representations using memory-guided training, and attenuates the predictive uncertainty of visual relations using a Gaussian Mixture Model (GMM). Extensive experiments demonstrate that our method achieves significant (up to 10% in some cases) performance gain over existing methods highlight- ing its superiority in generating more unbiased scene graphs. Code: https://github.com/sayaknag/unbiasedSGG.git
****
Dynamic Graph Learning With Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection
Yuan Wang, Kun Yu, Chen Chen, Xiyuan Hu, Silong Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7278-7287
With the springing up of face synthesis techniques, it is prominent in need to develop powerful face forgery detection methods due to security concerns. Some existing methods attempt to employ auxiliary frequency-aware information combined with CNN backbones to discover the forged clues. Due to the inadequate information interaction with image content, the extracted frequency features are thus spatially irrelavant, struggling to generalize well on increasingly realistic counterfeit types. To address this issue, we propose a Spatial-Frequency Dynamic Graph method to exploit the relation-aware features in spatial and frequency domains via dynamic graph learning. To this end, we introduce three well-designed components: 1) Content-guided Adaptive Frequency Extraction module to mine the content-adaptive forged frequency clues. 2) Multiple Domains Attention Map Learning module to enrich the spatial-frequency contextual features with multiscale attention maps. 3) Dynamic Graph Spatial-Frequency Feature Fusion Network to explore the high-order relation of spatial and frequency features. Extensive experiments on several benchmark show that our proposed method sustainedly exceeds the state-of-the-arts by a considerable margin.
****
Visual Language Pretrained Multiple Instance Zero-Shot Transfer for Histopathology Images
Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, Faisal Mahmood; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19764-19775
Contrastive visual language pretraining has emerged as a powerful method for either training new language-aware image encoders or augmenting existing pretrained models with zero-shot visual recognition capabilities. However, existing works typically train on large datasets of image-text pairs and have been designed to perform downstream tasks involving only small to medium sized-images, neither of which are applicable to the emerging field of computational pathology where there are limited publicly available paired image-text datasets and each image can span up to 100,000 x 100,000 pixels in dimensions. In this paper we present MI-Zero, a simple and intuitive framework for unleashing the zero-shot transfer capabilities of contrastively aligned image and text models to gigapixel histopathology whole slide images, enabling multiple downstream diagnostic tasks to be carried out by pretrained encoders without requiring any additional labels. MI-Zero

reformulates zero-shot transfer under the framework of multiple instance learnin
g to overcome the computational challenge of inference on extremely large images
. We used over 550k pathology reports and other available in-domain text corpora
 to pretrain our text encoder. By effectively leveraging strong pretrained encod
ers, our best model pretrained on over 33k histopathology image-caption pairs ac
hieves an average median zero-shot accuracy of 70.2% across three different real
-world cancer subtyping tasks. Our code is available at: https://github.com/mahm
oodlab/MI-Zero.
****
MIST: Multi-Modal Iterative Spatial-Temporal Transformer for Long-Form Video Que
stion Answering
Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, Mike Zheng Shou; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2023, pp. 14773-14783
To build Video Question Answering (VideoQA) systems capable of assisting humans
in daily activities, seeking answers from long-form videos with diverse and comp
lex events is a must. Existing multi-modal VQA models achieve promising performa
nce on images or short video clips, especially with the recent success of large-
scale multi-modal pre-training. However, when extending these methods to long-fo
rm videos, new challenges arise. On the one hand, using a dense video sampling s
trategy is computationally prohibitive. On the other hand, methods relying on sp
arse sampling struggle in scenarios where multi-event and multi-granularity visu
al reasoning are required. In this work, we introduce a new model named Multi-mo
dal Iterative Spatial-temporal Transformer (MIST) to better adapt pre-trained mo
dels for long-form VideoQA. Specifically, MIST decomposes traditional dense spat
ial-temporal self-attention into cascaded segment and region selection modules t
hat adaptively select frames and image regions that are closely relevant to the
question itself. Visual concepts at different granularities are then processed e
fficiently through an attention module. In addition, MIST iteratively conducts s
election and attention over multiple layers to support reasoning over multiple e
vents. The experimental results on four VideoQA datasets, including AGQA, NExT-Q
A, STAR, and Env-QA, show that MIST achieves state-of-the-art performance and is
 superior at computation efficiency and interpretability.
****
PMR: Prototypical Modal Rebalance for Multimodal Learning
Yunfeng Fan, Wenchao Xu, Haozhao Wang, Junxiao Wang, Song Guo; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 20029-20038
Multimodal learning (MML) aims to jointly exploit the common priors of different
 modalities to compensate for their inherent limitations. However, existing MML
methods often optimize a uniform objective for different modalities, leading to
the notorious "modality imbalance" problem and counterproductive MML performance
. To address the problem, some existing methods modulate the learning pace based
 on the fused modality, which is dominated by the better modality and eventually
 results in a limited improvement on the worse modal. To better exploit the feat
ures of multimodal, we propose Prototypical Modality Rebalance (PMR) to perform
stimulation on the particular slow-learning modality without interference from o
ther modalities. Specifically, we introduce the prototypes that represent genera
l features for each class, to build the non-parametric classifiers for uni-modal
 performance evaluation. Then, we try to accelerate the slow-learning modality b
y enhancing its clustering toward prototypes. Furthermore, to alleviate the supp
ression from the dominant modality, we introduce a prototype-based entropy regul
arization term during the early training stage to prevent premature convergence.
 Besides, our method only relies on the representations of each modality and wit
hout restrictions from model structures and fusion methods, making it with great
 application potential for various scenarios. The source code is available here.
****
Two-Stage Co-Segmentation Network Based on Discriminative Representation for Rec
overing Human Mesh From Videos
Boyang Zhang, Kehua Ma, Suping Wu, Zhixiang Yuan; Proceedings of the IEEE/CVF Co

nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5662-5670
Recovering 3D human mesh from videos has recently made significant progress. How ever, most of the existing methods focus on the temporal consistency of videos, while ignoring the spatial representation in complex scenes, thus failing to rec over a reasonable and smooth human mesh sequence under extreme illumination and chaotic backgrounds.To alleviate this problem, we propose a two-stage co-segment ation network based on discriminative representation for recovering human body m eshes from videos. Specifically, the first stage of the network segments the vid eo spatial domain to spotlight spatially fine-grained information, and then lear ns and enhances the intra-frame discriminative representation through a dual-exc itation mechanism and a frequency domain enhancement module, while suppressing i rrelevant information (e.g., background). The second stage focuses on temporal c ontext by segmenting the video temporal domain, and models inter-frame discrimin ative representation via a dynamic integration strategy.Further, to efficiently generate reasonable human discriminative actions, we carefully elaborate a landm ark anchor area loss to constrain the variation of the human motion area. Extens ive experimental results on large publicly available datasets indicate the super iority in comparison with most state-of-the-art. Code will be made public.
****
Multi-Sensor Large-Scale Dataset for Multi-View 3D Reconstruction
Oleg Voynov, Gleb Bobrovskikh, Pavel Karpyshev, Saveliy Galochkin, Andrei-Timote i Ardelean, Arseniy Bozhenko, Ekaterina Karmanova, Pavel Kopanev, Yaroslav Labut in-Rymsho, Ruslan Rakhimov, Aleksandr Safin, Valerii Serpiva, Alexey Artemov, Ev geny Burnaev, Dzmitry Tsetserukou, Denis Zorin; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21392-21403
We present a new multi-sensor dataset for multi-view 3D surface reconstruction. It includes registered RGB and depth data from sensors of different resolutions and modalities: smartphones, Intel RealSense, Microsoft Kinect, industrial camer as, and structured-light scanner. The scenes are selected to emphasize a diverse set of material properties challenging for existing algorithms. We provide arou nd 1.4 million images of 107 different scenes acquired from 100 viewing directio ns under 14 lighting conditions. We expect our dataset will be useful for evalua tion and training of 3D reconstruction algorithms and for related tasks. The dat aset is available at skoltech3d.appliedai.tech.
****
Privacy-Preserving Representations Are Not Enough: Recovering Scene Content From Camera Poses
Kunal Chelani, Torsten Sattler, Fredrik Kahl, Zuzana Kukelova; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 13132-13141
Visual localization is the task of estimating the camera pose from which a given image was taken and is central to several 3D computer vision applications. With the rapid growth in the popularity of AR/VR/MR devices and cloud-based applicat ions, privacy issues are becoming a very important aspect of the localization pr ocess. Existing work on privacy-preserving localization aims to defend against a n attacker who has access to a cloud-based service. In this paper, we show that an attacker can learn about details of a scene without any access by simply quer ying a localization service. The attack is based on the observation that modern visual localization algorithms are robust to variations in appearance and geomet ry. While this is in general a desired property, it also leads to algorithms loc alizing objects that are similar enough to those present in a scene. An attacker can thus query a server with a large enough set of images of objects, e.g., obt ained from the Internet, and some of them will be localized. The attacker can th us learn about object placements from the camera poses returned by the service (which is the minimal information returned by such a service). In this paper, we develop a proof-of-concept version of this attack and demonstrate its practical feasibility. The attack does not place any requirements on the localization algo rithm used, and thus also applies to privacy-preserving representations. Current work on privacy-preserving representations alone is thus insufficient.
****

Learning Anchor Transformations for 3D Garment Animation
Fang Zhao, Zekun Li, Shaoli Huang, Junwu Weng, Tianfei Zhou, Guo-Sen Xie, Jue Wang, Ying Shan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 491-500
This paper proposes an anchor-based deformation model, namely AnchorDEF, to predict 3D garment animation from a body motion sequence. It deforms a garment mesh template by a mixture of rigid transformations with extra nonlinear displacements. A set of anchors around the mesh surface is introduced to guide the learning of rigid transformation matrices. Once the anchor transformations are found, per-vertex nonlinear displacements of the garment template can be regressed in a canonical space, which reduces the complexity of deformation space learning. By explicitly constraining the transformed anchors to satisfy the consistencies of position, normal and direction, the physical meaning of learned anchor transformations in space is guaranteed for better generalization. Furthermore, an adaptive anchor updating is proposed to optimize the anchor position by being aware of local mesh topology for learning representative anchor transformations. Qualitative and quantitative experiments on different types of garments demonstrate that AnchorDEF achieves the state-of-the-art performance on 3D garment deformation prediction in motion, especially for loose-fitting garments.
****
Actionlet-Dependent Contrastive Learning for Unsupervised Skeleton-Based Action Recognition
Lilang Lin, Jiahang Zhang, Jiaying Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2363-2372
The self-supervised pretraining paradigm has achieved great success in skeleton-based action recognition. However, these methods treat the motion and static parts equally, and lack an adaptive design for different parts, which has a negative impact on the accuracy of action recognition. To realize the adaptive action modeling of both parts, we propose an Actionlet-Dependent Contrastive Learning method (ActCLR). The actionlet, defined as the discriminative subset of the human skeleton, effectively decomposes motion regions for better action modeling. In detail, by contrasting with the static anchor without motion, we extract the motion region of the skeleton data, which serves as the actionlet, in an unsupervised manner. Then, centering on actionlet, a motion-adaptive data transformation method is built. Different data transformations are applied to actionlet and non-actionlet regions to introduce more diversity while maintaining their own characteristics. Meanwhile, we propose a semantic-aware feature pooling method to build feature representations among motion and static regions in a distinguished manner. Extensive experiments on NTU RGB+D and PKUMMD show that the proposed method achieves remarkable action recognition performance. More visualization and quantitative experiments demonstrate the effectiveness of our method.
****
Ref-NPR: Reference-Based Non-Photorealistic Radiance Fields for Controllable Scene Stylization
Yuechen Zhang, Zexin He, Jinbo Xing, Xufeng Yao, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4242-4251
Current 3D scene stylization methods transfer textures and colors as styles using arbitrary style references, lacking meaningful semantic correspondences. We introduce Reference-Based Non-Photorealistic Radiance Fields (Ref-NPR) to address this limitation. This controllable method stylizes a 3D scene using radiance fields with a single stylized 2D view as a reference. We propose a ray registration process based on the stylized reference view to obtain pseudo-ray supervision in novel views. Then we exploit semantic correspondences in content images to fill occluded regions with perceptually similar styles, resulting in non-photorealistic and continuous novel view sequences. Our experimental results demonstrate that Ref-NPR outperforms existing scene and video stylization methods regarding visual quality and semantic correspondence. The code and data are publicly available on the project page at https://ref-npr.github.io.
****

PanoHead: Geometry-Aware 3D Full-Head Synthesis in 360deg
Sizhe An, Hongyi Xu, Yichun Shi, Guoxian Song, Umit Y. Ogras, Linjie Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20950-20959
Synthesis and reconstruction of 3D human head has gained increasing interests in computer vision and computer graphics recently. Existing state-of-the-art 3D generative adversarial networks (GANs) for 3D human head synthesis are either limited to near-frontal views or hard to preserve 3D consistency in large view angles. We propose PanoHead, the first 3D-aware generative model that enables high-quality view-consistent image synthesis of full heads in 360deg with diverse appearance and detailed geometry using only in-the-wild unstructured images for training. At its core, we lift up the representation power of recent 3D GANs and bridge the data alignment gap when training from in-the-wild images with widely distributed views. Specifically, we propose a novel two-stage self-adaptive image alignment for robust 3D GAN training. We further introduce a tri-grid neural volume representation that effectively addresses front-face and back-head feature entanglement rooted in the widely-adopted tri-plane formulation. Our method instills prior knowledge of 2D image segmentation in adversarial learning of 3D neural scene structures, enabling compositable head synthesis in diverse backgrounds. Benefiting from these designs, our method significantly outperforms previous 3D GANs, generating high-quality 3D heads with accurate geometry and diverse appearances, even with long wavy and afro hairstyles, renderable from arbitrary poses. Furthermore, we show that our system can reconstruct full 3D heads from single input images for personalized realistic 3D avatars.
****
Rethinking Feature-Based Knowledge Distillation for Face Recognition
Jingzhi Li, Zidong Guo, Hui Li, Seungju Han, Ji-won Baek, Min Yang, Ran Yang, Sungjoo Suh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20156-20165
With the continual expansion of face datasets, feature-based distillation prevails for large-scale face recognition. In this work, we attempt to remove identity supervision in student training, to spare the GPU memory from saving massive class centers. However, this naive removal leads to inferior distillation result. We carefully inspect the performance degradation from the perspective of intrinsic dimension, and argue that the gap in intrinsic dimension, namely the intrinsic gap, is intimately connected to the infamous capacity gap problem. By constraining the teacher's search space with reverse distillation, we narrow the intrinsic gap and unleash the potential of feature-only distillation. Remarkably, the proposed reverse distillation creates universally student-friendly teacher that demonstrates outstanding student improvement. We further enhance its effectiveness by designing a student proxy to better bridge the intrinsic gap. As a result, the proposed method surpasses state-of-the-art distillation techniques with identity supervision on various face recognition benchmarks, and the improvements are consistent across different teacher-student pairs.
****
NeurOCS: Neural NOCS Supervision for Monocular 3D Object Localization
Zhixiang Min, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Enrique Dunn, Manmohan Chandraker; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21404-21414
Monocular 3D object localization in driving scenes is a crucial task, but challenging due to its ill-posed nature. Estimating 3D coordinates for each pixel on the object surface holds great potential as it provides dense 2D-3D geometric constraints for the underlying PnP problem. However, high-quality ground truth supervision is not available in driving scenes due to sparsity and various artifacts of Lidar data, as well as the practical infeasibility of collecting per-instance CAD models. In this work, we present NeurOCS, a framework that uses instance masks and 3D boxes as input to learn 3D object shapes by means of differentiable rendering, which further serves as supervision for learning dense object coordinates. Our approach rests on insights in learning a category-level shape prior directly from real driving scenes, while properly handling single-view ambiguities

. Furthermore, we study and make critical design choices to learn object coordin ates more effectively from an object-centric view. Altogether, our framework lea ds to new state-of-the-art in monocular 3D localization that ranks 1st on the KI TTI-Object benchmark among published monocular methods.
****
Tree Instance Segmentation With Temporal Contour Graph
Adnan Firoze, Cameron Wingren, Raymond A. Yeh, Bedrich Benes, Daniel Aliaga; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2193-2202
We present a novel approach to perform instance segmentation, and counting, for densely packed self-similar trees using a top-view RGB image sequence. We propos e a solution that leverages pixel content, shape, and self-occlusion. First, we perform an initial over-segmentation of the image sequence and aggregate structu ral characteristics into a contour graph with temporal information incorporated. Second, using a graph convolutional network and its inherent local messaging pa ssing abilities, we merge adjacent tree crown patches into a final set of tree c rowns. Per various studies and comparisons, our method is superior to all prior methods and results in high-accuracy instance segmentation and counting, despite the trees being tightly packed. Finally, we provide various forest image sequen ce datasets suitable for subsequent benchmarking and evaluation captured at diff erent altitudes and leaf conditions.
****
A New Dataset Based on Images Taken by Blind People for Testing the Robustness o f Image Classification Models Trained for ImageNet Categories
Reza Akbarian Bafghi, Danna Gurari; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2023, pp. 16261-16270
Our goal is to improve upon the status quo for designing image classification mo dels trained in one domain that perform well on images from another domain. Comp lementing existing work in robustness testing, we introduce the first dataset fo r this purpose which comes from an authentic use case where photographers wanted to learn about the content in their images. We built a new test set using 8,900 images taken by people who are blind for which we collected metadata to indicat e the presence versus absence of 200 ImageNet object categories. We call this da taset VizWiz-Classification. We characterize this dataset and how it compares to the mainstream datasets for evaluating how well ImageNet-trained classification models generalize. Finally, we analyze the performance of 100 ImageNet classifi cation models on our new test dataset. Our fine-grained analysis demonstrates th at these models struggle on images with quality issues. To enable future extensi ons to this work, we share our new dataset with evaluation server at: https://vi zwiz.org/tasks-and-datasets/image-classification
****
Detecting Backdoors During the Inference Stage Based on Corruption Robustness Co nsistency
Xiaogeng Liu, Minghui Li, Haoyu Wang, Shengshan Hu, Dengpan Ye, Hai Jin, Libing Wu, Chaowei Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16363-16372
Deep neural networks are proven to be vulnerable to backdoor attacks. Detecting the trigger samples during the inference stage, i.e., the test-time trigger samp le detection, can prevent the backdoor from being triggered. However, existing d etection methods often require the defenders to have high accessibility to victi m models, extra clean data, or knowledge about the appearance of backdoor trigge rs, limiting their practicality. In this paper, we propose the test-time corrupt ion robustness consistency evaluation (TeCo), a novel test-time trigger sample d etection method that only needs the hard-label outputs of the victim models with out any extra information. Our journey begins with the intriguing observation th at the backdoor-infected models have similar performance across different image corruptions for the clean images, but perform discrepantly for the trigger sampl es. Based on this phenomenon, we design TeCo to evaluate test-time robustness co nsistency by calculating the deviation of severity that leads to predictions' tr ansition across different corruptions. Extensive experiments demonstrate that co

mpared with state-of-the-art defenses, which even require either certain informa tion about the trigger types or accessibility of clean data, TeCo outperforms th em on different backdoor attacks, datasets, and model architectures, enjoying a higher AUROC by 10% and 5 times of stability. The code is available at https://g ithub.com/CGCL-codes/TeCo
****
Black-Box Sparse Adversarial Attack via Multi-Objective Optimisation
Phoenix Neale Williams, Ke Li; Proceedings of the IEEE/CVF Conference on Compute r Vision and Pattern Recognition (CVPR), 2023, pp. 12291-12301
Deep neural networks (DNNs) are susceptible to adversarial images, raising conce rns about their reliability in safety-critical tasks. Sparse adversarial attacks , which limit the number of modified pixels, have shown to be highly effective i n causing DNNs to misclassify. However, existing methods often struggle to simul taneously minimize the number of modified pixels and the size of the modificatio ns, often requiring a large number of queries and assuming unrestricted access t o the targeted DNN. In contrast, other methods that limit the number of modified pixels often permit unbounded modifications, making them easily detectable. To address these limitations, we propose a novel multi-objective sparse attack algo rithm that efficiently minimizes the number of modified pixels and their size du ring the attack process. Our algorithm draws inspiration from evolutionary compu tation and incorporates a mechanism for prioritizing objectives that aligns with an attacker's goals. Our approach outperforms existing sparse attacks on CIFAR-10 and ImageNet trained DNN classifiers while requiring only a small query budge t, attaining competitive attack success rates while perturbing fewer pixels. Ove rall, our proposed attack algorithm provides a solution to the limitations of cu rrent sparse attack methods by jointly minimizing the number of modified pixels and their size. Our results demonstrate the effectiveness of our approach in res tricted scenarios, highlighting its potential to enhance DNN security.
****
Renderable Neural Radiance Map for Visual Navigation
Obin Kwon, Jeongho Park, Songhwai Oh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9099-9108
We propose a novel type of map for visual navigation, a renderable neural radian ce map (RNR-Map), which is designed to contain the overall visual information of a 3D environment. The RNR-Map has a grid form and consists of latent codes at e ach pixel. These latent codes are embedded from image observations, and can be c onverted to the neural radiance field which enables image rendering given a came ra pose. The recorded latent codes implicitly contain visual information about t he environment, which makes the RNR-Map visually descriptive. This visual inform ation in RNR-Map can be a useful guideline for visual localization and navigatio n. We develop localization and navigation frameworks that can effectively utiliz e the RNR-Map. We evaluate the proposed frameworks on camera tracking, visual lo calization, and image-goal navigation. Experimental results show that the RNR-Ma p-based localization framework can find the target location based on a single qu ery image with fast speed and competitive accuracy compared to other baselines. Also, this localization framework is robust to environmental changes, and even f inds the most visually similar places when a query image from a different enviro nment is given. The proposed navigation framework outperforms the existing image -goal navigation methods in difficult scenarios, under odometry and actuation no ises. The navigation framework shows 65.7% success rate in curved scenarios of t he NRNS dataset, which is an improvement of 18.6% over the current state-of-the-art. Project page: https://rllab-snu.github.io/projects/RNR-Map/
****
Revisiting Reverse Distillation for Anomaly Detection
Tran Dinh Tien, Anh Tuan Nguyen, Nguyen Hoang Tran, Ta Duc Huy, Soan T.M. Duong, Chanh D. Tr. Nguyen, Steven Q. H. Truong; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24511-24520
Anomaly detection is an important application in large-scale industrial manufact uring. Recent methods for this task have demonstrated excellent accuracy but com e with a latency trade-off. Memory based approaches with dominant performances l

ike PatchCore or Coupled-hypersphere-based Feature Adaptation (CFA) require an e
xternal memory bank, which significantly lengthens the execution time. Another a
pproach that employs Reversed Distillation (RD) can perform well while maintaini
ng low latency. In this paper, we revisit this idea to improve its performance,
establishing a new state-of-the-art benchmark on the challenging MVTec dataset f
or both anomaly detection and localization. The proposed method, called RD++, ru
ns six times faster than PatchCore, and two times faster than CFA but introduces
a negligible latency compared to RD. We also experiment on the BTAD and Retinal
OCT datasets to demonstrate our method's generalizability and conduct important
ablation experiments to provide insights into its configurations. Source code w
ill be available at https://github.com/tientrandinh/Revisiting-Reverse-Distillat
ion.
****
Diffusion-Based Generation, Optimization, and Planning in 3D Scenes
Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang
, Song-Chun Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2023, pp. 16750-16761
We introduce SceneDiffuser, a conditional generative model for 3D scene understa
nding. SceneDiffuser provides a unified model for solving scene-conditioned gene
ration, optimization, and planning. In contrast to prior works, SceneDiffuser is
intrinsically scene-aware, physics-based, and goal-oriented. With an iterative
sampling strategy, SceneDiffuser jointly formulates the scene-aware generation,
physics-based optimization, and goal-oriented planning via a diffusion-based den
oising process in a fully differentiable fashion. Such a design alleviates the d
iscrepancies among different modules and the posterior collapse of previous scen
e-conditioned generative models. We evaluate SceneDiffuser with various 3D scene
understanding tasks, including human pose and motion generation, dexterous gras
p generation, path planning for 3D navigation, and motion planning for robot arm
s. The results show significant improvements compared with previous models, demo
nstrating the tremendous potential of SceneDiffuser for the broad community of 3
D scene understanding.
****
TMO: Textured Mesh Acquisition of Objects With a Mobile Device by Using Differen
tiable Rendering
Jaehoon Choi, Dongki Jung, Taejae Lee, Sangwook Kim, Youngdong Jung, Dinesh Mano
cha, Donghwan Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2023, pp. 16674-16684
We present a new pipeline for acquiring a textured mesh in the wild with a singl
e smartphone which offers access to images, depth maps, and valid poses. Our met
hod first introduces an RGBD-aided structure from motion, which can yield filter
ed depth maps and refines camera poses guided by corresponding depth. Then, we a
dopt the neural implicit surface reconstruction method, which allows for high qu
ality mesh and develops a new training process for applying a regularization pro
vided by classical multi-view stereo methods. Moreover, we apply a differentiabl
e rendering to fine-tune incomplete texture maps and generate textures which are
perceptually closer to the original scene. Our pipeline can be applied to any c
ommon objects in the real world without the need for either in-the-lab environme
nts or accurate mask images. We demonstrate results of captured objects with com
plex shapes and validate our method numerically against existing 3D reconstructi
on and texture mapping methods.
****
Meta-Causal Learning for Single Domain Generalization
Jin Chen, Zhi Gao, Xinxiao Wu, Jiebo Luo; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7683-7692
Single domain generalization aims to learn a model from a single training domain
(source domain) and apply it to multiple unseen test domains (target domains).
Existing methods focus on expanding the distribution of the training domain to c
over the target domains, but without estimating the domain shift between the sou
rce and target domains. In this paper, we propose a new learning paradigm, namel
y simulate-analyze-reduce, which first simulates the domain shift by building an

auxiliary domain as the target domain, then learns to analyze the causes of domain shift, and finally learns to reduce the domain shift for model adaptation. Under this paradigm, we propose a meta-causal learning method to learn meta-knowledge, that is, how to infer the causes of domain shift between the auxiliary and source domains during training. We use the meta-knowledge to analyze the shift between the target and source domains during testing. Specifically, we perform multiple transformations on source data to generate the auxiliary domain, perform counterfactual inference to learn to discover the causal factors of the shift between the auxiliary and source domains, and incorporate the inferred causality into factor-aware domain alignments. Extensive experiments on several benchmarks of image classification show the effectiveness of our method.
****
Grad-PU: Arbitrary-Scale Point Cloud Upsampling via Gradient Descent With Learned Distance Functions
Yun He, Danhang Tang, Yinda Zhang, Xiangyang Xue, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5354-5363
Most existing point cloud upsampling methods have roughly three steps: feature extraction, feature expansion and 3D coordinate prediction. However, they usually suffer from two critical issues: (1) fixed upsampling rate after one-time training, since the feature expansion unit is customized for each upsampling rate; (2) outliers or shrinkage artifact caused by the difficulty of precisely predicting 3D coordinates or residuals of upsampled points. To adress them, we propose a new framework for accurate point cloud upsampling that supports arbitrary upsampling rates. Our method first interpolates the low-res point cloud according to a given upsampling rate. And then refine the positions of the interpolated points with an iterative optimization process, guided by a trained model estimating the difference between the current point cloud and the high-res target. Extensive quantitative and qualitative results on benchmarks and downstream tasks demonstrate that our method achieves the state-of-the-art accuracy and efficiency.
****
Trainable Projected Gradient Method for Robust Fine-Tuning
Junjiao Tian, Zecheng He, Xiaoliang Dai, Chih-Yao Ma, Yen-Cheng Liu, Zsolt Kira; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7836-7845
Recent studies on transfer learning have shown that selectively fine-tuning a subset of layers or customizing different learning rates for each layer can greatly improve robustness to out-of-distribution (OOD) data and retain generalization capability in the pre-trained models. However, most of these methods employ manually crafted heuristics or expensive hyper-parameter search, which prevent them from scaling up to large datasets and neural networks. To solve this problem, we propose Trainable Projected Gradient Method (TPGM) to automatically learn the constraint imposed for each layer for a fine-grained fine-tuning regularization. This is motivated by formulating fine-tuning as a bi-level constrained optimization problem. Specifically, TPGM maintains a set of projection radii, i.e., distance constraints between the fine-tuned model and the pre-trained model, for each layer, and enforces them through weight projections. To learn the constraints, we propose a bi-level optimization to automatically learn the best set of projection radii in an end-to-end manner. Theoretically, we show that the bi-level optimization formulation is the key to learn different constraints for each layer. Empirically, with little hyper-parameter search cost, TPGM outperforms existing fine-tuning methods in OOD performance while matching the best in-distribution (ID) performance. For example, when fine-tuned on DomainNet-Real and ImageNet, compared to vanilla fine-tuning, TPGM shows 22% and 10% relative OOD improvement respectively on their sketch counterparts.
****
Text2Scene: Text-Driven Indoor Scene Stylization With Part-Aware Details
Inwoo Hwang, Hyeonwoo Kim, Young Min Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1890-1899
We propose Text2Scene, a method to automatically create realistic textures for v

irtual scenes composed of multiple objects. Guided by a reference image and text descriptions, our pipeline adds detailed texture on labeled 3D geometries in the room such that the generated colors respect the hierarchical structure or semantic parts that are often composed of similar materials. Instead of applying flat stylization on the entire scene at a single step, we obtain weak semantic cues from geometric segmentation, which are further clarified by assigning initial colors to segmented parts. Then we add texture details for individual objects such that their projections on image space exhibit feature embedding aligned with the embedding of the input. The decomposition makes the entire pipeline tractable to a moderate amount of computation resources and memory. As our framework utilizes the existing resources of image and text embedding, it does not require dedicated datasets with high-quality textures designed by skillful artists. To the best of our knowledge, it is the first practical and scalable approach that can create detailed and realistic textures of the desired style that maintain structural context for scenes with multiple objects.
****
FEND: A Future Enhanced Distribution-Aware Contrastive Learning Framework for Long-Tail Trajectory Prediction
Yuning Wang, Pu Zhang, Lei Bai, Jianru Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1400-1409
Predicting the future trajectories of the traffic agents is a gordian technique in autonomous driving. However, trajectory prediction suffers from data imbalance in the prevalent datasets, and the tailed data is often more complicated and safety-critical. In this paper, we focus on dealing with the long-tail phenomenon in trajectory prediction. Previous methods dealing with long-tail data did not take into account the variety of motion patterns in the tailed data. In this paper, we put forward a future enhanced contrastive learning framework to recognize tail trajectory patterns and form a feature space with separate pattern clusters.Furthermore, a distribution aware hyper predictor is brought up to better utilize the shaped feature space.Our method is a model-agnostic framework and can be plugged into many well-known baselines. Experimental results show that our framework outperforms the state-of-the-art long-tail prediction method on tailed samples by 9.5% on ADE and 8.5% on FDE, while maintaining or slightly improving the averaged performance. Our method also surpasses many long-tail techniques on trajectory prediction task.
****
MP-Former: Mask-Piloted Transformer for Image Segmentation
Hao Zhang, Feng Li, Huaizhe Xu, Shijia Huang, Shilong Liu, Lionel M. Ni, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18074-18083
We present a mask-piloted Transformer which improves masked-attention in Mask2Former for image segmentation. The improvement is based on our observation that Mask2Former suffers from inconsistent mask predictions between consecutive decoder layers, which leads to inconsistent optimization goals and low utilization of decoder queries. To address this problem, we propose a mask-piloted training approach, which additionally feeds noised ground-truth masks in masked-attention and trains the model to reconstruct the original ones. Compared with the predicted masks used in mask-attention, the ground-truth masks serve as a pilot and effectively alleviate the negative impact of inaccurate mask predictions in Mask2Former. Based on this technique, our MP-Former achieves a remarkable performance improvement on all three image segmentation tasks (instance, panoptic, and semantic), yielding +2.3 AP and +1.6 mIoU on the Cityscapes instance and semantic segmentation tasks with a ResNet-50 backbone. Our method also significantly speeds up the training, outperforming Mask2Former with half of the number of training epochs on ADE20K with both a ResNet-50 and a Swin-L backbones. Moreover, our method only introduces little computation during training and no extra computation during inference. Our code will be released at https://github.com/IDEA-Research/MP-Former.
****
HDR Imaging With Spatially Varying Signal-to-Noise Ratios

Yiheng Chi, Xingguang Zhang, Stanley H. Chan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5724-5734
While today's high dynamic range (HDR) image fusion algorithms are capable of blending multiple exposures, the acquisition is often controlled so that the dynamic range within one exposure is narrow. For HDR imaging in photon-limited situations, the dynamic range can be enormous and the noise within one exposure is spatially varying. Existing image denoising algorithms and HDR fusion algorithms both fail to handle this situation, leading to severe limitations in low-light HDR imaging. This paper presents two contributions. Firstly, we identify the source of the problem. We find that the issue is associated with the co-existence of (1) spatially varying signal-to-noise ratio, especially the excessive noise due to very dark regions, and (2) a wide luminance range within each exposure. We show that while the issue can be handled by a bank of denoisers, the complexity is high. Secondly, we propose a new method called the spatially varying high dynamic range (SV-HDR) fusion network to simultaneously denoise and fuse images. We introduce a new exposure-shared block within our custom-designed multi-scale transformer framework. In a variety of testing conditions, the performance of the proposed SV-HDR is better than the existing methods.
****
Learning Orthogonal Prototypes for Generalized Few-Shot Semantic Segmentation
Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, Ting Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11319-11328
Generalized few-shot semantic segmentation (GFSS) distinguishes pixels of base and novel classes from the background simultaneously, conditioning on sufficient data of base classes and a few examples from novel class. A typical GFSS approach has two training phases: base class learning and novel class updating. Nevertheless, such a stand-alone updating process often compromises the well-learnt features and results in performance drop on base classes. In this paper, we propose a new idea of leveraging Projection onto Orthogonal Prototypes (POP), which updates features to identify novel classes without compromising base classes. POP builds a set of orthogonal prototypes, each of which represents a semantic class, and makes the prediction for each class separately based on the features projected onto its prototype. Technically, POP first learns prototypes on base data, and then extends the prototype set to novel classes. The orthogonal constraint of POP encourages the orthogonality between the learnt prototypes and thus mitigates the influence on base class features when generalizing to novel prototypes. Moreover, we capitalize on the residual of feature projection as the background representation to dynamically fit semantic shifting (i.e., background no longer includes the pixels of novel classes in updating phase). Extensive experiments on two benchmarks demonstrate that our POP achieves superior performances on novel classes without sacrificing much accuracy on base classes. Notably, POP outperforms the state-of-the-art fine-tuning by 3.93% overall mIoU on PASCAL-5i in 5-shot scenario.
****
TAPS3D: Text-Guided 3D Textured Shape Generation From Pseudo Supervision
Jiacheng Wei, Hao Wang, Jiashi Feng, Guosheng Lin, Kim-Hui Yap; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16805-16815
In this paper, we investigate an open research task of generating controllable 3D textured shapes from the given textual descriptions. Previous works either require ground truth caption labeling or extensive optimization time. To resolve these issues, we present a novel framework, TAPS3D, to train a text-guided 3D shape generator with pseudo captions. Specifically, based on rendered 2D images, we retrieve relevant words from the CLIP vocabulary and construct pseudo captions using templates. Our constructed captions provide high-level semantic supervision for generated 3D shapes. Further, in order to produce fine-grained textures and increase geometry diversity, we propose to adopt low-level image regularization to enable fake-rendered images to align with the real ones. During the inference phase, our proposed model can generate 3D textured shapes from the given text

without any additional optimization. We conduct extensive experiments to analyze each of our proposed components and show the efficacy of our framework in generating high-fidelity 3D textured and text-relevant shapes.
****
Are Deep Neural Networks SMARTer Than Second Graders?
Anoop Cherian, Kuan-Chuan Peng, Suhas Lohit, Kevin A. Smith, Joshua B. Tenenbaum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10834-10844
Recent times have witnessed an increasing number of applications of deep neural networks towards solving tasks that require superior cognitive abilities, e.g., playing Go, generating art, question answering (such as ChatGPT), etc. Such a dramatic progress raises the question: how generalizable are neural networks in solving problems that demand broad skills? To answer this question, we propose SMART: a Simple Multimodal Algorithmic Reasoning Task and the associated SMART-101 dataset, for evaluating the abstraction, deduction, and generalization abilities of neural networks in solving visuo-linguistic puzzles designed specifically for children in the 6--8 age group. Our dataset consists of 101 unique puzzles; each puzzle comprises a picture and a question, and their solution needs a mix of several elementary skills, including arithmetic, algebra, and spatial reasoning, among others. To scale our dataset towards training deep neural networks, we programmatically generate entirely new instances for each puzzle while retaining their solution algorithm. To benchmark the performance on the SMART-101 dataset, we propose a vision-and-language meta-learning model that can incorporate varied state-of-the-art neural backbones. Our experiments reveal that while powerful deep models offer reasonable performances on puzzles in a supervised setting, they are not better than random accuracy when analyzed for generalization -- filling this gap may demand new multimodal learning approaches.
****
Reliability in Semantic Segmentation: Are We on the Right Track?
Pau de Jorge, Riccardo Volpi, Philip H.S. Torr, Grégory Rogez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7173-7182
Motivated by the increasing popularity of transformers in computer vision, in recent times there has been a rapid development of novel architectures. While in-domain performance follows a constant, upward trend, properties like robustness or uncertainty estimation are less explored -leaving doubts about advances in model reliability. Studies along these axes exist, but they are mainly limited to classification models. In contrast, we carry out a study on semantic segmentation, a relevant task for many real-world applications where model reliability is paramount. We analyze a broad variety of models, spanning from older ResNet-based architectures to novel transformers and assess their reliability based on four metrics: robustness, calibration, misclassification detection and out-of-distribution (OOD) detection. We find that while recent models are significantly more robust, they are not overall more reliable in terms of uncertainty estimation. We further explore methods that can come to the rescue and show that improving calibration can also help with other uncertainty metrics such as misclassification or OOD detection. This is the first study on modern segmentation models focused on both robustness and uncertainty estimation and we hope it will help practitioners and researchers interested in this fundamental vision task.
****
Video Test-Time Adaptation for Action Recognition
Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, Horst Bischof; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22952-22961
Although action recognition systems can achieve top performance when evaluated on in-distribution test points, they are vulnerable to unanticipated distribution shifts in test data. However, test-time adaptation of video action recognition models against common distribution shifts has so far not been demonstrated. We propose to address this problem with an approach tailored to spatio-temporal models that is capable of adaptation on a single video sample at a step. It consists

in a feature distribution alignment technique that aligns online estimates of t
est set statistics towards the training statistics. We further enforce predictio
n consistency over temporally augmented views of the same test video sample. Eva
luations on three benchmark action recognition datasets show that our proposed t
echnique is architecture-agnostic and able to significantly boost the performanc
e on both, the state of the art convolutional architecture TANet and the Video S
win Transformer. Our proposed method demonstrates a substantial performance gain
 over existing test-time adaptation approaches in both evaluations of a single d
istribution shift and the challenging case of random distribution shifts.
****
Bi-Level Meta-Learning for Few-Shot Domain Generalization
Xiaorong Qin, Xinhang Song, Shuqiang Jiang; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15900-15910
The goal of few-shot learning is to learn the generalizability from seen to unse
en data with only a few samples. Most previous few-shot learning focus on learni
ng generalizability within particular domains. However, the more practical scena
rios may also require generalizability across domains. In this paper, we study t
he problem of Few-shot domain generalization (FSDG), which is a more challenging
 variant of few-shot classification. FSDG requires additional generalization wit
h larger gap from seen domains to unseen domains. We address FSDG problem by met
a-learning two levels of meta-knowledge, where the lower-level meta-knowledge ar
e domain-specific embedding spaces as subspaces of a base space for intra-domain
 generalization, and the upper-level meta-knowledge is the base space and a prio
r subspace over domain-specific spaces for inter-domain generalization. We formu
late the two levels of meta-knowledge learning problem with bi-level optimizatio
n, and further develop an optimization algorithm without Hessian information to
solve it. We demonstrate our method is significantly superior to the previous wo
rks by evaluating it on the widely used benchmark Meta-Dataset.
****
Tensor4D: Efficient Neural 4D Decomposition for High-Fidelity Dynamic Reconstruc
tion and Rendering
Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, Yebin Liu; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2023, pp. 16632-16642
We present Tensor4D, an efficient yet effective approach to dynamic scene modeli
ng. The key of our solution is an efficient 4D tensor decomposition method so th
at the dynamic scene can be directly represented as a 4D spatio-temporal tensor.
 To tackle the accompanying memory issue, we decompose the 4D tensor hierarchica
lly by projecting it first into three time-aware volumes and then nine compact f
eature planes. In this way, spatial information over time can be simultaneously
captured in a compact and memory-efficient manner. When applying Tensor4D for dy
namic scene reconstruction and rendering, we further factorize the 4D fields to
different scales in the sense that structural motions and dynamic detailed chang
es can be learned from coarse to fine. The effectiveness of our method is valida
ted on both synthetic and real-world scenes. Extensive experiments show that our
 method is able to achieve high-quality dynamic reconstruction and rendering fro
m sparse-view camera rigs or even a monocular camera.
****
Blowing in the Wind: CycleNet for Human Cinemagraphs From Still Images
Hugo Bertiche, Niloy J. Mitra, Kuldeep Kulkarni, Chun-Hao P. Huang, Tuanfeng Y.
Wang, Meysam Madadi, Sergio Escalera, Duygu Ceylan; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 459-468
Cinemagraphs are short looping videos created by adding subtle motions to a stat
ic image. This kind of media is popular and engaging. However, automatic generat
ion of cinemagraphs is an underexplored area and current solutions require tedio
us low-level manual authoring by artists. In this paper, we present an automatic
 method that allows generating human cinemagraphs from single RGB images. We inv
estigate the problem in the context of dressed humans under the wind. At the cor
e of our method is a novel cyclic neural network that produces looping cinemagra
phs for the target loop duration. To circumvent the problem of collecting real d

ata, we demonstrate that it is possible, by working in the image normal space, to learn garment motion dynamics on synthetic data and generalize to real data. We evaluate our method on both synthetic and real data and demonstrate that it is possible to create compelling and plausible cinemagraphs from single RGB images.
****
Learning Personalized High Quality Volumetric Head Avatars From Monocular RGB Videos
Ziqian Bai, Feitong Tan, Zeng Huang, Kripasindhu Sarkar, Danhang Tang, Di Qiu, Abhimitra Meka, Ruofei Du, Mingsong Dou, Sergio Orts-Escolano, Rohit Pandey, Ping Tan, Thabo Beeler, Sean Fanello, Yinda Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16890-16900
We propose a method to learn a high-quality implicit 3D head avatar from a monocular RGB video captured in the wild. The learnt avatar is driven by a parametric face model to achieve user-controlled facial expressions and head poses. Our hybrid pipeline combines the geometry prior and dynamic tracking of a 3DMM with a neural radiance field to achieve fine-grained control and photorealism. To reduce over-smoothing and improve out-of-model expressions synthesis, we propose to predict local features anchored on the 3DMM geometry. These learnt features are driven by 3DMM deformation and interpolated in 3D space to yield the volumetric radiance at a designated query point. We further show that using a Convolutional Neural Network in the UV space is critical in incorporating spatial context and producing representative local features. Extensive experiments show that we are able to reconstruct high-quality avatars, with more accurate expression-dependent details, good generalization to out-of-training expressions, and quantitatively superior renderings compared to other state-of-the-art approaches.
****
Multi-Modal Learning With Missing Modality via Shared-Specific Feature Modelling
Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, Gustavo Carneiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15878-15887
The missing modality issue is critical but non-trivial to be solved by multi-modal models. Current methods aiming to handle the missing modality problem in multi-modal tasks, either deal with missing modalities only during evaluation or train separate models to handle specific missing modality settings. In addition, these models are designed for specific tasks, so for example, classification models are not easily adapted to segmentation tasks and vice versa. In this paper, we propose the Shared-Specific Feature Modelling (ShaSpec) method that is considerably simpler and more effective than competing approaches that address the issues above. ShaSpec is designed to take advantage of all available input modalities during training and evaluation by learning shared and specific features to better represent the input data. This is achieved from a strategy that relies on auxiliary tasks based on distribution alignment and domain classification, in addition to a residual feature fusion procedure. Also, the design simplicity of ShaSpec enables its easy adaptation to multiple tasks, such as classification and segmentation. Experiments are conducted on both medical image segmentation and computer vision classification, with results indicating that ShaSpec outperforms competing methods by a large margin. For instance, on BraTS2018, ShaSpec improves the SOTA by more than 3% for enhancing tumour, 5% for tumour core and 3% for whole tumour.
****
Panoptic Compositional Feature Field for Editable Scene Rendering With Network-Inferred Labels via Metric Learning
Xinhua Cheng, Yanmin Wu, Mengxi Jia, Qian Wang, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4947-4957
Despite neural implicit representations demonstrating impressive high-quality view synthesis capacity, decomposing such representations into objects for instance-level editing is still challenging. Recent works learn object-compositional representations supervised by ground truth instance annotations and produce promis

ing scene editing results. However, ground truth annotations are manually labeled and expensive in practice, which limits their usage in real-world scenes. In this work, we attempt to learn an object-compositional neural implicit representation for editable scene rendering by leveraging labels inferred from the off-the-shelf 2D panoptic segmentation networks instead of the ground truth annotations. We propose a novel framework named Panoptic Compositional Feature Field (PCFF), which introduces an instance quadruplet metric learning to build a discriminating panoptic feature space for reliable scene editing. In addition, we propose semantic-related strategies to further exploit the correlations between semantic and appearance attributes for achieving better rendering results. Experiments on multiple scene datasets including ScanNet, Replica, and ToyDesk demonstrate that our proposed method achieves superior performance for novel view synthesis and produces convincing real-world scene editing results. The code will be available.
****
Progressive Backdoor Erasing via Connecting Backdoor and Adversarial Attacks
Bingxu Mu, Zhenxing Niu, Le Wang, Xue Wang, Qiguang Miao, Rong Jin, Gang Hua; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20495-20503
Deep neural networks (DNNs) are known to be vulnerable to both backdoor attacks as well as adversarial attacks. In the literature, these two types of attacks are commonly treated as distinct problems and solved separately, since they belong to training-time and inference-time attacks respectively. However, in this paper we find an intriguing connection between them: for a model planted with backdoors, we observe that its adversarial examples have similar behaviors as its triggered samples, i.e., both activate the same subset of DNN neurons. It indicates that planting a backdoor into a model will significantly affect the model's adversarial examples. Based on this observations, a novel Progressive Backdoor Erasing (PBE) algorithm is proposed to progressively purify the infected model by leveraging untargeted adversarial attacks. Different from previous backdoor defense methods, one significant advantage of our approach is that it can erase backdoor even when the additional clean dataset is unavailable. We empirically show that, against 5 state-of-the-art backdoor attacks, our AFT can effectively erase the backdoor triggers without obvious performance degradation on clean samples and significantly outperforms existing defense methods.
****
LayoutFormer++: Conditional Graphic Layout Generation via Constraint Serialization and Decoding Space Restriction
Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, Dongmei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18403-18412
Conditional graphic layout generation, which generates realistic layouts according to user constraints, is a challenging task that has not been well-studied yet. First, there is limited discussion about how to handle diverse user constraints flexibly and uniformly. Second, to make the layouts conform to user constraints, existing work often sacrifices generation quality significantly. In this work, we propose LayoutFormer++ to tackle the above problems. First, to flexibly handle diverse constraints, we propose a constraint serialization scheme, which represents different user constraints as sequences of tokens with a predefined format. Then, we formulate conditional layout generation as a sequence-to-sequence transformation, and leverage encoder-decoder framework with Transformer as the basic architecture. Furthermore, to make the layout better meet user requirements without harming quality, we propose a decoding space restriction strategy. Specifically, we prune the predicted distribution by ignoring the options that definitely violate user constraints and likely result in low-quality layouts, and make the model samples from the restricted distribution. Experiments demonstrate that LayoutFormer++ outperforms existing approaches on all the tasks in terms of both better generation quality and less constraint violation.
****

DisWOT: Student Architecture Search for Distillation WithOut Training
Peijie Dong, Lujun Li, Zimian Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11898-11908
Knowledge distillation (KD) is an effective training strategy to improve the lightweight student models under the guidance of cumbersome teachers. However, the large architecture difference across the teacher-student pairs limits the distillation gains. In contrast to previous adaptive distillation methods to reduce the teacher-student gap, we explore a novel training-free framework to search for the best student architectures for a given teacher. Our work first empirically show that the optimal model under vanilla training cannot be the winner in distillation. Secondly, we find that the similarity of feature semantics and sample relations between random-initialized teacher-student networks have good correlations with final distillation performances. Thus, we efficiently measure similarity matrixs conditioned on the semantic activation maps to select the optimal student via an evolutionary algorithm without any training. In this way, our student architecture search for Distillation WithOut Training (DisWOT) significantly improves the performance of the model in the distillation stage with at least 180x training acceleration. Additionally, we extend similarity metrics in DisWOT as new distillers and KD-based zero-proxies. Our experiments on CIFAR, ImageNet and NAS-Bench-201 demonstrate that our technique achieves state-of-the-art results on different search spaces. Our project and code are available at https://lilujunai.github.io/DisWOT-CVPR2023/.
****
Stare at What You See: Masked Image Modeling Without Reconstruction
Hongwei Xue, Peng Gao, Hongyang Li, Yu Qiao, Hao Sun, Houqiang Li, Jiebo Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22732-22741
Masked Autoencoders (MAE) have been prevailing paradigms for large-scale vision representation pre-training. By reconstructing masked image patches from a small portion of visible image regions, MAE forces the model to infer semantic correlation within an image. Recently, some approaches apply semantic-rich teacher models to extract image features as the reconstruction target, leading to better performance. However, unlike the low-level features such as pixel values, we argue the features extracted by powerful teacher models already encode rich semantic correlation across regions in an intact image. This raises one question: is reconstruction necessary in Masked Image Modeling (MIM) with a teacher model? In this paper, we propose an efficient MIM paradigm named MaskAlign. MaskAlign simply learns the consistency of visible patch feature extracted by the student model and intact image features extracted by the teacher model. To further advance the performance and tackle the problem of input inconsistency between the student and teacher model, we propose a Dynamic Alignment (DA) module to apply learnable alignment. Our experimental results demonstrate that masked modeling does not lose effectiveness even without reconstruction on masked regions. Combined with Dynamic Alignment, MaskAlign can achieve state-of-the-art performance with much higher efficiency.
****
Joint Visual Grounding and Tracking With Natural Language Specification
Li Zhou, Zikun Zhou, Kaige Mao, Zhenyu He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23151-23160
Tracking by natural language specification aims to locate the referred target in a sequence based on the natural language description. Existing algorithms solve this issue in two steps, visual grounding and tracking, and accordingly deploy the separated grounding model and tracking model to implement these two steps, respectively. Such a separated framework overlooks the link between visual grounding and tracking, which is that the natural language descriptions provide global semantic cues for localizing the target for both two steps. Besides, the separated framework can hardly be trained end-to-end. To handle these issues, we propose a joint visual grounding and tracking framework, which reformulates grounding and tracking as a unified task: localizing the referred target based on the given visual-language references. Specifically, we propose a multi-source relation

modeling module to effectively build the relation between the visual-language re ferences and the test image. In addition, we design a temporal modeling module t o provide a temporal clue with the guidance of the global semantic information f or our model, which effectively improves the adaptability to the appearance vari ations of the target. Extensive experimental results on TNL2K, LaSOT, OTB99, and RefCOCOg demonstrate that our method performs favorably against state-of-the-ar t algorithms for both tracking and grounding. Code is available at https://githu b.com/lizhou-cs/JointNLT.
****
Neural Kaleidoscopic Space Sculpting
Byeongjoo Ahn, Michael De Zeeuw, Ioannis Gkioulekas, Aswin C. Sankaranarayanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio n (CVPR), 2023, pp. 4349-4358
We introduce a method that recovers full-surround 3D reconstructions from a sing le kaleidoscopic image using a neural surface representation. Full-surround 3D r econstruction is critical for many applications, such as augmented and virtual r eality. A kaleidoscope, which uses a single camera and multiple mirrors, is a co nvenient way of achieving full-surround coverage, as it redistributes light dire ctions and thus captures multiple viewpoints in a single image. This enables sin gle-shot and dynamic full-surround 3D reconstruction. However, using a kaleidosc opic image for multi-view stereo is challenging, as we need to decompose the ima ge into multi-view images by identifying which pixel corresponds to which virtua l camera, a process we call labeling. To address this challenge, pur approach av oids the need to explicitly estimate labels, but instead "sculpts" a neural surf ace representation through the careful use of silhouette, background, foreground , and texture information present in the kaleidoscopic image. We demonstrate the advantages of our method in a range of simulated and real experiments, on both static and dynamic scenes.
****
Few-Shot Semantic Image Synthesis With Class Affinity Transfer
Marlène Careil, Jakob Verbeek, Stéphane Lathuilière; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23611-2 3620
Semantic image synthesis aims to generate photo realistic images given a semanti c segmentation map. Despite much recent progress, training them still requires l arge datasets of images annotated with per-pixel label maps that are extremely t edious to obtain. To alleviate the high annotation cost, we propose a transfer m ethod that leverages a model trained on a large source dataset to improve the le arning ability on small target datasets via estimated pairwise relations between source and target classes. The class affinity matrix is introduced as a first l ayer to the source model to make it compatible with the target label maps, and t he source model is then further fine-tuned for the target domain. To estimate th e class affinities we consider different approaches to leverage prior knowledge: semantic segmentation on the source domain, textual label embeddings, and self- supervised vision features. We apply our approach to GAN-based and diffusion-bas ed architectures for semantic synthesis. Our experiments show that the different ways to estimate class affinity can effectively combined, and that our approach significantly improves over existing state-of-the-art transfer approaches for g enerative image models.
****
Implicit Identity Driven Deepfake Face Swapping Detection
Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, Dengpan Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2023, pp. 4490-4499
In this paper, we consider the face swapping detection from the perspective of f ace identity. Face swapping aims to replace the target face with the source face and generate the fake face that the human cannot distinguish between real and f ake. We argue that the fake face contains the explicit identity and implicit ide ntity, which respectively corresponds to the identity of the source face and tar get face during face swapping. Note that the explicit identities of faces can be

extracted by regular face recognizers. Particularly, the implicit identity of real face is consistent with the its explicit identity. Thus the difference between explicit and implicit identity of face facilitates face swapping detection. Following this idea, we propose a novel implicit identity driven framework for face swapping detection. Specifically, we design an explicit identity contrast (EIC) loss and an implicit identity exploration (IIE) loss, which supervises a CNN backbone to embed face images into the implicit identity space. Under the guidance of EIC, real samples are pulled closer to their explicit identities, while fake samples are pushed away from their explicit identities. Moreover, IIE is derived from the margin-based classification loss function, which encourages the fake faces with known target identities to enjoy intra-class compactness and inter-class diversity. Extensive experiments and visualizations on several datasets demonstrate the generalization of our method against the state-of-the-art counterparts.
****
Class Relationship Embedded Learning for Source-Free Unsupervised Domain Adaptation
Yixin Zhang, Zilei Wang, Weinan He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7619-7629
This work focuses on a practical knowledge transfer task defined as Source-Free Unsupervised Domain Adaptation (SFUDA), where only a well-trained source model and unlabeled target data are available. To fully utilize source knowledge, we propose to transfer the class relationship, which is domain-invariant but still under-explored in previous works. To this end, we first regard the classifier weights of the source model as class prototypes to compute class relationship, and then propose a novel probability-based similarity between target-domain samples by embedding the source-domain class relationship, resulting in Class Relationship embedded Similarity (CRS). Here the inter-class term is particularly considered in order to more accurately represent the similarity between two samples, in which the source prior of class relationship is utilized by weighting. Finally, we propose to embed CRS into contrastive learning in a unified form. Here both class-aware and instance discrimination contrastive losses are employed, which are complementary to each other. We combine the proposed method with existing representative methods to evaluate its efficacy in multiple SFUDA settings. Extensive experimental results reveal that our method can achieve state-of-the-art performance due to the transfer of domain-invariant class relationship.
****
Logical Consistency and Greater Descriptive Power for Facial Hair Attribute Learning
Haiyu Wu, Grace Bezold, Aman Bhatta, Kevin W. Bowyer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8588-8597
Face attribute research has so far used only simple binary attributes for facial hair; e.g., beard / no beard. We have created a new, more descriptive facial hair annotation scheme and applied it to create a new facial hair attribute dataset, FH37K. Face attribute research also so far has not dealt with logical consistency and completeness. For example, in prior research, an image might be classified as both having no beard and also having a goatee (a type of beard). We show that the test accuracy of previous classification methods on facial hair attribute classification drops significantly if logical consistency of classifications is enforced. We propose a logically consistent prediction loss, LCPLoss, to aid learning of logical consistency across attributes, and also a label compensation training strategy to eliminate the problem of no positive prediction across a set of related attributes. Using an attribute classifier trained on FH37K, we investigate how facial hair affects face recognition accuracy, including variation across demographics. Results show that similarity and difference in facial hairstyle have important effects on the impostor and genuine score distributions in face recognition. The code is at https:// github.com/ HaiyuWu/ facial hair logical.
****

One-to-Few Label Assignment for End-to-End Dense Detection
Shuai Li, Minghan Li, Ruihuang Li, Chenhang He, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7350-7359
One-to-one (o2o) label assignment plays a key role for transformer based end-to-end detection, and it has been recently introduced in fully convolutional detectors for lightweight end-to-end dense detection. However, o2o can largely degrade the feature learning performance due to the limited number of positive samples. Though extra positive samples can be introduced to mitigate this issue, the computation of self- and cross- attentions among anchors prevents its practical application to dense and fully convolutional detectors. In this work, we propose a simple yet effective one-to-few (o2f) label assignment strategy for end-to-end dense detection. Apart from defining one positive and many negative anchors for each object, we define several soft anchors, which serve as positive and negative samples simultaneously. The positive and negative weights of these soft anchors are dynamically adjusted during training so that they can contribute more to 'representation learning' in the early training stage and contribute more to 'duplicated prediction removal' in the later stage. The detector trained in this way can not only learn a strong feature representation but also perform end-to-end detection. Experiments on COCO and CrowdHuman datasets demonstrate the effectiveness of the proposed o2f scheme.
****
Spatio-Temporal Pixel-Level Contrastive Learning-Based Source-Free Domain Adaptation for Video Semantic Segmentation
Shao-Yuan Lo, Poojan Oza, Sumanth Chennupati, Alejandro Galindo, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10534-10543
Unsupervised Domain Adaptation (UDA) of semantic segmentation transfers labeled source knowledge to an unlabeled target domain by relying on accessing both the source and target data. However, the access to source data is often restricted or infeasible in real-world scenarios. Under the source data restrictive circumstances, UDA is less practical. To address this, recent works have explored solutions under the Source-Free Domain Adaptation (SFDA) setup, which aims to adapt a source-trained model to the target domain without accessing source data. Still, existing SFDA approaches use only image-level information for adaptation, making them sub-optimal in video applications. This paper studies SFDA for Video Semantic Segmentation (VSS), where temporal information is leveraged to address video adaptation. Specifically, we propose Spatio-Temporal Pixel-Level (STPL) contrastive learning, a novel method that takes full advantage of spatio-temporal information to tackle the absence of source data better. STPL explicitly learns semantic correlations among pixels in the spatio-temporal space, providing strong self-supervision for adaptation to the unlabeled target domain. Extensive experiments show that STPL achieves state-of-the-art performance on VSS benchmarks compared to current UDA and SFDA approaches. Code is available at: https://github.com/shaoyuanlo/STPL
****
InternImage: Exploring Large-Scale Vision Foundation Models With Deformable Convolutions
Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14408-14419
Compared to the great progress of large-scale vision transformers (ViTs) in recent years, large-scale models based on convolutional neural networks (CNNs) are still in an early state. This work presents a new large-scale CNN-based foundation model, termed InternImage, which can obtain the gain from increasing parameters and training data like ViTs. Different from the recent CNNs that focus on large dense kernels, InternImage takes deformable convolution as the core operator, so that our model not only has the large effective receptive field required for downstream tasks such as detection and segmentation, but also has the adaptive s

patial aggregation conditioned by input and task information. As a result, the p
roposed InternImage reduces the strict inductive bias of traditional CNNs and ma
kes it possible to learn stronger and more robust patterns with large-scale para
meters from massive data like ViTs. The effectiveness of our model is proven on
challenging benchmarks including ImageNet, COCO, and ADE20K. It is worth mention
ing that InternImage-H achieved a new record 65.4 mAP on COCO test-dev and 62.9
mIoU on ADE20K, outperforming current leading CNNs and ViTs.
****
DAA: A Delta Age AdaIN Operation for Age Estimation via Binary Code Transformer
Ping Chen, Xingpeng Zhang, Ye Li, Ju Tao, Bin Xiao, Bing Wang, Zongjie Jiang; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2023, pp. 15836-15845
Naked eye recognition of age is usually based on comparison with the age of othe
rs. However, this idea is ignored by computer tasks because it is difficult to o
btain representative contrast images of each age. Inspired by the transfer learn
ing, we designed the Delta Age AdaIN (DAA) operation to obtain the feature diffe
rence with each age, which obtains the style map of each age through the learned
 values representing the mean and standard deviation. We let the input of transf
er learning as the binary code of age natural number to obtain continuous age fe
ature information. The learned two groups of values in Binary code mapping are c
orresponding to the mean and standard deviation of the comparison ages. In summa
ry, our method consists of four parts: FaceEncoder, DAA operation, Binary code m
apping, and AgeDecoder modules. After getting the delta age via AgeDecoder, we t
ake the average value of all comparison ages and delta ages as the predicted age
. Compared with state-of-the-art methods, our method achieves better performance
 with fewer parameters on multiple facial age datasets. Code is available at htt
ps://github.com/redcping/Delta_Age_AdaIN
****
Fake It Till You Make It: Learning Transferable Representations From Synthetic I
mageNet Clones
Mert Bülent Sar■y■ld■z, Karteek Alahari, Diane Larlus, Yannis Kalantidis; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 8011-8021
Recent image generation models such as Stable Diffusion have exhibited an impres
sive ability to generate fairly realistic images starting from a simple text pro
mpt. Could such models render real images obsolete for training image prediction
 models? In this paper, we answer part of this provocative question by investiga
ting the need for real images when training models for ImageNet classification.
Provided only with the class names that have been used to build the dataset, we
explore the ability of Stable Diffusion to generate synthetic clones of ImageNet
 and measure how useful these are for training classification models from scratc
h. We show that with minimal and class-agnostic prompt engineering, ImageNet clo
nes are able to close a large part of the gap between models produced by synthet
ic images and models trained with real images, for the several standard classifi
cation benchmarks that we consider in this study. More importantly, we show that
 models trained on synthetic images exhibit strong generalization properties and
 perform on par with models trained on real data for transfer. Project page: htt
ps://europe.naverlabs.com/imagenet-sd
****
Mind the Label Shift of Augmentation-Based Graph OOD Generalization
Junchi Yu, Jian Liang, Ran He; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2023, pp. 11620-11630
Out-of-distribution (OOD) generalization is an important issue for Graph Neural
Networks (GNNs). Recent works employ different graph editions to generate augmen
ted environments and learn an invariant GNN for generalization. However, the gra
ph structural edition inevitably alters the graph label. This causes the label s
hift in augmentations and brings inconsistent predictive relationships among aug
mented environments. To address this issue, we propose LiSA, which generates lab
el-invariant augmentations to facilitate graph OOD generalization. Instead of re
sorting to graph editions, LiSA exploits Label-invariant Subgraphs of the traini

ng graphs to construct Augmented environments. Specifically, LiSA first designs the variational subgraph generators to efficiently extract locally predictive patterns and construct multiple label-invariant subgraphs. Then, the subgraphs produced by different generators are collected to build different augmented environments. To promote diversity among augmented environments, LiSA further introduces a tractable energy-based regularization to enlarge pair-wise distances between the distributions of environments. In this manner, LiSA generates diverse augmented environments with a consistent predictive relationship to facilitate learning an invariant GNN. Extensive experiments on node-level and graph-level OOD benchmarks show that LiSA achieves impressive generalization performance with different GNN backbones. Code is available on https://github.com/Samyu0304/LiSA.
****
Unsupervised Intrinsic Image Decomposition With LiDAR Intensity
Shogo Sato, Yasuhiro Yao, Taiga Yoshida, Takuhiro Kaneko, Shingo Ando, Jun Shimamura; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13466-13475
Intrinsic image decomposition (IID) is the task that decomposes a natural image into albedo and shade. While IID is typically solved through supervised learning methods, it is not ideal due to the difficulty in observing ground truth albedo and shade in general scenes. Conversely, unsupervised learning methods are currently underperforming supervised learning methods since there are no criteria for solving the ill-posed problems. Recently, light detection and ranging (LiDAR) is widely used due to its ability to make highly precise distance measurements. Thus, we have focused on the utilization of LiDAR, especially LiDAR intensity, to address this issue. In this paper, we propose unsupervised intrinsic image decomposition with LiDAR intensity (IID-LI). Since the conventional unsupervised learning methods consist of image-to-image transformations, simply inputting LiDAR intensity is not an effective approach. Therefore, we design an intensity consistency loss that computes the error between LiDAR intensity and gray-scaled albedo to provide a criterion for the ill-posed problem. In addition, LiDAR intensity is difficult to handle due to its sparsity and occlusion, hence, a LiDAR intensity densification module is proposed. We verified the estimating quality using our own dataset, which include RGB images, LiDAR intensity and human judged annotations. As a result, we achieved an estimation accuracy that outperforms conventional unsupervised learning methods.
****
HIER: Metric Learning Beyond Class Labels via Hierarchical Regularization
Sungyeon Kim, Boseung Jeong, Suha Kwak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19903-19912
Supervision for metric learning has long been given in the form of equivalence between human-labeled classes. Although this type of supervision has been a basis of metric learning for decades, we argue that it hinders further advances in the field. In this regard, we propose a new regularization method, dubbed HIER, to discover the latent semantic hierarchy of training data, and to deploy the hierarchy to provide richer and more fine-grained supervision than inter-class separability induced by common metric learning losses. HIER achieves this goal with no annotation for the semantic hierarchy but by learning hierarchical proxies in hyperbolic spaces. The hierarchical proxies are learnable parameters, and each of them is trained to serve as an ancestor of a group of data or other proxies to approximate the semantic hierarchy among them. HIER deals with the proxies along with data in hyperbolic space since the geometric properties of the space are well-suited to represent their hierarchical structure. The efficacy of HIER is evaluated on four standard benchmarks, where it consistently improved the performance of conventional methods when integrated with them, and consequently achieved the best records, surpassing even the existing hyperbolic metric learning technique, in almost all settings.
****
Diffusion Probabilistic Model Made Slim
Xingyi Yang, Daquan Zhou, Jiashi Feng, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22552-2

Despite the visually-pleasing results achieved, the massive computational cost has been a long-standing flaw for diffusion probabilistic models (DPMs), which, in turn, greatly limits their applications on resource-limited platforms. Prior methods towards efficient DPM, however, have largely focused on accelerating the testing yet overlooked their huge complexity and size. In this paper, we make a dedicated attempt to lighten DPM while striving to preserve its favourable performance. We start by training a small-sized latent diffusion model (LDM) from scratch but observe a significant fidelity drop in the synthetic images. Through a thorough assessment, we find that DPM is intrinsically biased against high-frequency generation, and learns to recover different frequency components at different time-steps. These properties make compact networks unable to represent frequency dynamics with accurate high-frequency estimation. Towards this end, we introduce a customized design for slim DPM, which we term as Spectral Diffusion (SD), for lightweight image synthesis. SD incorporates wavelet gating in its architecture to enable frequency dynamic feature extraction at every reverse steps, and conducts spectrum-aware distillation to promote high-frequency recovery by inverse weighting the objective based on spectrum magnitudes. Experimental results demonstrate that, SD achieves 8-18x computational complexity reduction as compared to the latent diffusion models on a series of conditional and unconditional image generation tasks while retaining competitive image fidelity.
****
Confidence-Aware Personalized Federated Learning via Variational Expectation Maximization
Junyi Zhu, Xingchen Ma, Matthew B. Blaschko; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24542-24551
Federated Learning (FL) is a distributed learning scheme to train a shared model across clients. One common and fundamental challenge in FL is that the sets of data across clients could be non-identically distributed and have different sizes. Personalized Federated Learning (PFL) attempts to solve this challenge via locally adapted models. In this work, we present a novel framework for PFL based on hierarchical Bayesian modeling and variational inference. A global model is introduced as a latent variable to augment the joint distribution of clients' parameters and capture the common trends of different clients, optimization is derived based on the principle of maximizing the marginal likelihood and conducted using variational expectation maximization. Our algorithm gives rise to a closed-form estimation of a confidence value which comprises the uncertainty of clients' parameters and local model deviations from the global model. The confidence value is used to weigh clients' parameters in the aggregation stage and adjust the regularization effect of the global model. We evaluate our method through extensive empirical studies on multiple datasets. Experimental results show that our approach obtains competitive results under mild heterogeneous circumstances while significantly outperforming state-of-the-art PFL frameworks in highly heterogeneous settings.
****
Hierarchical Supervision and Shuffle Data Augmentation for 3D Semi-Supervised Object Detection
Chuandong Liu, Chenqiang Gao, Fangcen Liu, Pengcheng Li, Deyu Meng, Xinbo Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23819-23828
State-of-the-art 3D object detectors are usually trained on large-scale datasets with high-quality 3D annotations. However, such 3D annotations are often expensive and time-consuming, which may not be practical for real applications. A natural remedy is to adopt semi-supervised learning (SSL) by leveraging a limited amount of labeled samples and abundant unlabeled samples. Current pseudo-labeling-based SSL object detection methods mainly adopt a teacher-student framework, with a single fixed threshold strategy to generate supervision signals, which inevitably brings confused supervision when guiding the student network training. Besides, the data augmentation of the point cloud in the typical teacher-student framework is too weak, and only contains basic down sampling and flip-and-shift (i

.e., rotate and scaling), which hinders the effective learning of feature information. Hence, we address these issues by introducing a novel approach of Hierarchical Supervision and Shuffle Data Augmentation (HSSDA), which is a simple yet effective teacher-student framework. The teacher network generates more reasonable supervision for the student network by designing a dynamic dual-threshold strategy. Besides, the shuffle data augmentation strategy is designed to strengthen the feature representation ability of the student network. Extensive experiments show that HSSDA consistently outperforms the recent state-of-the-art methods on different datasets. The code will be released at https://github.com/azhuantou/HSSDA.
****
Interactive and Explainable Region-Guided Radiology Report Generation
Tim Tanida, Philip Müller, Georgios Kaissis, Daniel Rueckert; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7433-7442
The automatic generation of radiology reports has the potential to assist radiologists in the time-consuming task of report writing. Existing methods generate the full report from image-level features, failing to explicitly focus on anatomical regions in the image. We propose a simple yet effective region-guided report generation model that detects anatomical regions and then describes individual, salient regions to form the final report. While previous methods generate reports without the possibility of human intervention and with limited explainability, our method opens up novel clinical use cases through additional interactive capabilities and introduces a high degree of transparency and explainability. Comprehensive experiments demonstrate our method's effectiveness in report generation, outperforming previous state-of-the-art models, and highlight its interactive capabilities. The code and checkpoints are available at https://github.com/ttanida/rgrg.
****
MED-VT: Multiscale Encoder-Decoder Video Transformer With Application To Object Segmentation
Rezaul Karim, He Zhao, Richard P. Wildes, Mennatullah Siam; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6323-6333
Multiscale video transformers have been explored in a wide variety of vision tasks. To date, however, the multiscale processing has been confined to the encoder or decoder alone. We present a unified multiscale encoder-decoder transformer that is focused on dense prediction tasks in videos. Multiscale representation at both encoder and decoder yields key benefits of implicit extraction of spatiotemporal features (i.e. without reliance on input optical flow) as well as temporal consistency at encoding and coarse-to-fine detection for high-level (e.g. object) semantics to guide precise localization at decoding. Moreover, we propose a transductive learning scheme through many-to-many label propagation to provide temporally consistent predictions.We showcase our Multiscale Encoder-Decoder Video Transformer (MED-VT) on Automatic Video Object Segmentation (AVOS) and actor/action segmentation, where we outperform state-of-the-art approaches on multiple benchmarks using only raw images, without using optical flow.
****
PET-NeuS: Positional Encoding Tri-Planes for Neural Surfaces
Yiqun Wang, Ivan Skorokhodov, Peter Wonka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12598-12607
A signed distance function (SDF) parametrized by an MLP is a common ingredient of neural surface reconstruction. We build on the successful recent method NeuS to extend it by three new components. The first component is to borrow the tri-plane representation from EG3D and represent signed distance fields as a mixture of tri-planes and MLPs instead of representing it with MLPs only. Using tri-planes leads to a more expressive data structure but will also introduce noise in the reconstructed surface. The second component is to use a new type of positional encoding with learnable weights to combat noise in the reconstruction process. We divide the features in the tri-plane into multiple frequency scales and modula

te them with sin and cos functions of different frequencies. The third component is to use learnable convolution operations on the tri-plane features using self-attention convolution to produce features with different frequency bands. The experiments show that PET-NeuS achieves high-fidelity surface reconstruction on standard datasets. Following previous work and using the Chamfer metric as the most important way to measure surface reconstruction quality, we are able to improve upon the NeuS baseline by 57% on Nerf-synthetic (0.84 compared to 1.97) and by 15.5% on DTU (0.71 compared to 0.84). The qualitative evaluation reveals how our method can better control the interference of high-frequency noise.
****
ZegCLIP: Towards Adapting CLIP for Zero-Shot Semantic Segmentation
Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, Yifan Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11175-11185
Recently, CLIP has been applied to pixel-level zero-shot learning tasks via a wo-stage scheme. The general idea is to first generate class-agnostic region proposals and then feed the cropped proposal regions to CLIP to utilize its image-level zero-shot classification capability. While effective, such a scheme requires two image encoders, one for proposal generation and one for CLIP, leading to a complicated pipeline and high computational cost. In this work, we pursue a simpler-and-efficient one-stage solution that directly extends CLIP's zero-shot prediction capability from image to pixel level. Our investigation starts with a straightforward extension as our baseline that generates semantic masks by comparing the similarity between text and patch embeddings extracted from CLIP. However, such a paradigm could heavily overfit the seen classes and fail to generalize to unseen classes. To handle this issue, we propose three simple-but-effective designs and figure out that they can significantly retain the inherent zero-shot capacity of CLIP and improve pixel-level generalization ability. Incorporating those modifications leads to an efficient zero-shot semantic segmentation system called ZegCLIP. Through extensive experiments on three public benchmarks, ZegCLIP demonstrates superior performance, outperforming the state-of-the-art methods by a large margin under both "inductive" and "transductive" zero-shot settings. In addition, compared with the two-stage method, our one-stage ZegCLIP achieves a speedup of about 5 times faster during inference. We release the code at https://github.com/ZiqinZhou66/ZegCLIP.git.
****
AdaptiveMix: Improving GAN Training via Feature Space Shrinkage
Haozhe Liu, Wentian Zhang, Bing Li, Haoqian Wu, Nanjun He, Yawen Huang, Yuexiang Li, Bernard Ghanem, Yefeng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16219-16229
Due to the outstanding capability for data generation, Generative Adversarial Networks (GANs) have attracted considerable attention in unsupervised learning. However, training GANs is difficult, since the training distribution is dynamic for the discriminator, leading to unstable image representation. In this paper, we address the problem of training GANs from a novel perspective, i.e., robust image classification. Motivated by studies on robust image representation, we propose a simple yet effective module, namely AdaptiveMix, for GANs, which shrinks the regions of training data in the image representation space of the discriminator. Considering it is intractable to directly bound feature space, we propose to construct hard samples and narrow down the feature distance between hard and easy samples. The hard samples are constructed by mixing a pair of training images. We evaluate the effectiveness of our AdaptiveMix with widely-used and state-of-the-art GAN architectures. The evaluation results demonstrate that our AdaptiveMix can facilitate the training of GANs and effectively improve the image quality of generated samples. We also show that our AdaptiveMix can be further applied to image classification and Out-Of-Distribution (OOD) detection tasks, by equipping it with state-of-the-art methods. Extensive experiments on seven publicly available datasets show that our method effectively boosts the performance of baselines. The code is publicly available at https://github.com/WentianZhang-ML/AdaptiveMix.

****
Specialist Diffusion: Plug-and-Play Sample-Efficient Fine-Tuning of Text-to-Imag
e Diffusion Models To Learn Any Unseen Style
Haoming Lu, Hazarapet Tunanyan, Kai Wang, Shant Navasardyan, Zhangyang Wang, Hum
phrey Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
 Recognition (CVPR), 2023, pp. 14267-14276
Diffusion models have demonstrated impressive capability of text-conditioned ima
ge synthesis, and broader application horizons are emerging by personalizing tho
se pretrained diffusion models toward generating some specialized target object
or style. In this paper, we aim to learn an unseen style by simply fine-tuning a
 pre-trained diffusion model with a handful of images (e.g., less than 10), so t
hat the fine-tuned model can generate high-quality images of arbitrary objects i
n this style. Such extremely lowshot fine-tuning is accomplished by a novel tool
kit of finetuning techniques, including text-to-image customized data augmentati
ons, a content loss to facilitate content-style disentanglement, and sparse upda
ting that focuses on only a few time steps. Our framework, dubbed Specialist Dif
fusion, is plug-and-play to existing diffusion model backbones and other persona
lization techniques. We demonstrate it to outperform the latest few-shot persona
lization alternatives of diffusion models such as Textual Inversion and DreamBoo
th, in terms of learning highly sophisticated styles with ultra-sample-efficient
 tuning. We further show that Specialist Diffusion can be integrated on top of t
extual inversion to boost performance further, even on highly unusual styles. Ou
r codes are available at: https://github.com/Picsart-AI-Research/Specialist-Diff
usion
****
Benchmarking Self-Supervised Learning on Diverse Pathology Datasets
Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo, Sérgio Pereira; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
23, pp. 3344-3354
Computational pathology can lead to saving human lives, but models are annotatio
n hungry and pathology images are notoriously expensive to annotate. Self-superv
ised learning has shown to be an effective method for utilizing unlabeled data,
and its application to pathology could greatly benefit its downstream tasks. Yet
, there are no principled studies that compare SSL methods and discuss how to ad
apt them for pathology. To address this need, we execute the largest-scale study
 of SSL pre-training on pathology image data, to date. Our study is conducted us
ing 4 representative SSL methods on diverse downstream tasks. We establish that
large-scale domain-aligned pre-training in pathology consistently out-performs I
mageNet pre-training in standard SSL settings such as linear and fine-tuning eva
luations, as well as in low-label regimes. Moreover, we propose a set of domain-
specific techniques that we experimentally show leads to a performance boost. La
stly, for the first time, we apply SSL to the challenging task of nuclei instanc
e segmentation and show large and consistent performance improvements under dive
rse settings.
****
Planning-Oriented Autonomous Driving
Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, S
enyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Da
i, Yu Qiao, Hongyang Li; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2023, pp. 17853-17862
Modern autonomous driving system is characterized as modular tasks in sequential
 order, i.e., perception, prediction, and planning. In order to perform a wide d
iversity of tasks and achieve advanced-level intelligence, contemporary approach
es either deploy standalone models for individual tasks, or design a multi-task
paradigm with separate heads. However, they might suffer from accumulative error
s or deficient task coordination. Instead, we argue that a favorable framework s
hould be devised and optimized in pursuit of the ultimate goal, i.e., planning o
f the self-driving car. Oriented at this, we revisit the key components within p
erception and prediction, and prioritize the tasks such that all these tasks con
tribute to planning. We introduce Unified Autonomous Driving (UniAD), a comprehe

nsive framework up-to-date that incorporates full-stack driving tasks in one network. It is exquisitely devised to leverage advantages of each module, and provide complementary feature abstractions for agent interaction from a global perspective. Tasks are communicated with unified query interfaces to facilitate each other toward planning. We instantiate UniAD on the challenging nuScenes benchmark. With extensive ablations, the effectiveness of using such a philosophy is proven by substantially outperforming previous state-of-the-arts in all aspects. Code and models are public.
****
HyperCUT: Video Sequence From a Single Blurry Image Using Unsupervised Ordering
Bang-Dang Pham, Phong Tran, Anh Tran, Cuong Pham, Rang Nguyen, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9843-9852
We consider the challenging task of training models for image-to-video deblurring, which aims to recover a sequence of sharp images corresponding to a given blurry image input. A critical issue disturbing the training of an image-to-video model is the ambiguity of the frame ordering since both the forward and backward sequences are plausible solutions. This paper proposes an effective self-supervised ordering scheme that allows training high-quality image-to-video deblurring models. Unlike previous methods that rely on order-invariant losses, we assign an explicit order for each video sequence, thus avoiding the order-ambiguity issue. Specifically, we map each video sequence to a vector in a latent high-dimensional space so that there exists a hyperplane such that for every video sequence, the vectors extracted from it and its reversed sequence are on different sides of the hyperplane. The side of the vectors will be used to define the order of the corresponding sequence. Last but not least, we propose a real-image dataset for the image-to-video deblurring problem that covers a variety of popular domains, including face, hand, and street. Extensive experimental results confirm the effectiveness of our method. Code and data are available at https://github.com/VinAIResearch/HyperCUT.git
****
Can't Steal? Cont-Steal! Contrastive Stealing Attacks Against Image Encoders
Zeyang Sha, Xinlei He, Ning Yu, Michael Backes, Yang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16373-16383
Self-supervised representation learning techniques have been developing rapidly to make full use of unlabeled images. They encode images into rich features that are oblivious to downstream tasks. Behind their revolutionary representation power, the requirements for dedicated model designs and a massive amount of computation resources expose image encoders to the risks of potential model stealing attacks - a cheap way to mimic the well-trained encoder performance while circumventing the demanding requirements. Yet conventional attacks only target supervised classifiers given their predicted labels and/or posteriors, which leaves the vulnerability of unsupervised encoders unexplored. In this paper, we first instantiate the conventional stealing attacks against encoders and demonstrate their severer vulnerability compared with downstream classifiers. To better leverage the rich representation of encoders, we further propose Cont-Steal, a contrastive-learning-based attack, and validate its improved stealing effectiveness in various experiment settings. As a takeaway, we appeal to our community's attention to the intellectual property protection of representation learning techniques, especially to the defenses against encoder stealing attacks like ours.
****
Document Image Shadow Removal Guided by Color-Aware Background
Ling Zhang, Yinghao He, Qing Zhang, Zheng Liu, Xiaolong Zhang, Chunxia Xiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1818-1827
Existing works on document image shadow removal mostly depend on learning and leveraging a constant background (the color of the paper) from the image. However, the constant background is less representative and frequently ignores other background colors, such as the printed colors, resulting in distorted results. In t

his paper, we present a color-aware background extraction network (CBENet) for extracting a spatially varying background image that accurately depicts the background colors of the document. Furthermore, we propose a background-guided document images shadow removal network (BGShadowNet) using the predicted spatially varying background as auxiliary information, which consists of two stages. At Stage I, a background-constrained decoder is designed to promote a coarse result. Then, the coarse result is refined with a background-based attention module (BAModule) to maintain a consistent appearance and a detail improvement module (DEModule) to enhance the texture details at Stage II. Experiments on two benchmark datasets qualitatively and quantitatively validate the superiority of the proposed approach over state-of-the-arts.
****
Independent Component Alignment for Multi-Task Learning
Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, Anton Konushin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20083-20093
In a multi-task learning (MTL) setting, a single model is trained to tackle a diverse set of tasks jointly. Despite rapid progress in the field, MTL remains challenging due to optimization issues such as conflicting and dominating gradients. In this work, we propose using a condition number of a linear system of gradients as a stability criterion of an MTL optimization. We theoretically demonstrate that a condition number reflects the aforementioned optimization issues. Accordingly, we present Aligned-MTL, a novel MTL optimization approach based on the proposed criterion, that eliminates instability in the training process by aligning the orthogonal components of the linear system of gradients. While many recent MTL approaches guarantee convergence to a minimum, task trade-offs cannot be specified in advance. In contrast, Aligned-MTL provably converges to an optimal point with pre-defined task-specific weights, which provides more control over the optimization result. Through experiments, we show that the proposed approach consistently improves performance on a diverse set of MTL benchmarks, including semantic and instance segmentation, depth estimation, surface normal estimation, and reinforcement learning.
****
Edges to Shapes to Concepts: Adversarial Augmentation for Robust Vision
Aditay Tripathi, Rishubh Singh, Anirban Chakraborty, Pradeep Shenoy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24470-24479
Recent work has shown that deep vision models tend to be overly dependent on low-level or "texture" features, leading to poor generalization. Various data augmentation strategies have been proposed to overcome this so-called texture bias in DNNs. We propose a simple, lightweight adversarial augmentation technique that explicitly incentivizes the network to learn holistic shapes for accurate prediction in an object classification setting. Our augmentations superpose edgemaps from one image onto another image with shuffled patches, using a randomly determined mixing proportion, with the image label of the edgemap image. To classify these augmented images, the model needs to not only detect and focus on edges but distinguish between relevant and spurious edges. We show that our augmentations significantly improve classification accuracy and robustness measures on a range of datasets and neural architectures. As an example, for ViT-S, We obtain absolute gains on classification accuracy gains up to 6%. We also obtain gains of up to 28% and 8.5% on natural adversarial and out-of-distribution datasets like ImageNet-A (for ViTB) and ImageNet-R (for ViT-S), respectively. Analysis using a range of probe datasets shows substantially increased shape sensitivity in our trained models, explaining the observed improvement in robustness and classification accuracy.
****
ReVISE: Self-Supervised Speech Resynthesis With Visual Input for Universal and Generalized Speech Regeneration
Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, Yossi Adi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.

18795-18805
Prior works on improving speech quality with visual input typically study each type of auditory distortion separately (e.g., separation, inpainting, video-to-speech) and present tailored algorithms. This paper proposes to unify these subjects and study Generalized Speech Regeneration, where the goal is not to reconstruct the exact reference clean signal, but to focus on improving certain aspects of speech while not necessarily preserving the rest such as voice. In particular, this paper concerns intelligibility, quality, and video synchronization. We cast the problem as audio-visual speech resynthesis, which is composed of two steps: pseudo audio-visual speech recognition (P-AVSR) and pseudo text-to-speech synthesis (P-TTS). P-AVSR and P-TTS are connected by discrete units derived from a self-supervised speech model. Moreover, we utilize self-supervised audio-visual speech model to initialize P-AVSR. The proposed model is coined ReVISE. ReVISE is the first high-quality model for in-the-wild video-to-speech synthesis and achieves superior performance on all LRS3 audio-visual regeneration tasks with a single model. To demonstrates its applicability in the real world, ReVISE is also evaluated on EasyCom, an audio-visual benchmark collected under challenging acoustic conditions with only 1.6 hours of training data. Similarly, ReVISE greatly suppresses noise and improves quality. Project page: https://wnhsu.github.io/ReVISE.
****
Improved Distribution Matching for Dataset Condensation
Ganlong Zhao, Guanbin Li, Yipeng Qin, Yizhou Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7856-7865
Dataset Condensation aims to condense a large dataset into a smaller one while maintaining its ability to train a well-performing model, thus reducing the storage cost and training effort in deep learning applications. However, conventional dataset condensation methods are optimization-oriented and condense the dataset by performing gradient or parameter matching during model optimization, which is computationally intensive even on small datasets and models. In this paper, we propose a novel dataset condensation method based on distribution matching, which is more efficient and promising. Specifically, we identify two important shortcomings of naive distribution matching (i.e., imbalanced feature numbers and unvalidated embeddings for distance computation) and address them with three novel techniques (i.e., partitioning and expansion augmentation, efficient and enriched model sampling, and class-aware distribution regularization). Our simple yet effective method outperforms most previous optimization-oriented methods with much fewer computational resources, thereby scaling data condensation to larger datasets and models. Extensive experiments demonstrate the effectiveness of our method. Codes are available at https://github.com/uitrbn/IDM
****
Feature Separation and Recalibration for Adversarial Robustness
Woo Jae Kim, Yoonki Cho, Junsik Jung, Sung-Eui Yoon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8183-8192
Deep neural networks are susceptible to adversarial attacks due to the accumulation of perturbations in the feature level, and numerous works have boosted model robustness by deactivating the non-robust feature activations that cause model mispredictions. However, we claim that these malicious activations still contain discriminative cues and that with recalibration, they can capture additional useful information for correct model predictions. To this end, we propose a novel, easy-to-plugin approach named Feature Separation and Recalibration (FSR) that recalibrates the malicious, non-robust activations for more robust feature maps through Separation and Recalibration. The Separation part disentangles the input feature map into the robust feature with activations that help the model make correct predictions and the non-robust feature with activations that are responsible for model mispredictions upon adversarial attack. The Recalibration part then adjusts the non-robust activations to restore the potentially useful cues for model predictions. Extensive experiments verify the superiority of FSR compared to traditional deactivation techniques and demonstrate that it improves the robus

tness of existing adversarial training methods by up to 8.57% with small computa tional overhead. Codes are available at https://github.com/wkim97/FSR.
****
Nerflets: Local Radiance Fields for Efficient Structure-Aware 3D Scene Represent ation From 2D Supervision
Xiaoshuai Zhang, Abhijit Kundu, Thomas Funkhouser, Leonidas Guibas, Hao Su, Kyle Genova; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2023, pp. 8274-8284
We address efficient and structure-aware 3D scene representation from images. Ne rflets are our key contribution-- a set of local neural radiance fields that tog ether represent a scene. Each nerflet maintains its own spatial position, orient ation, and extent, within which it contributes to panoptic, density, and radianc e reconstructions. By leveraging only photometric and inferred panoptic image su pervision, we can directly and jointly optimize the parameters of a set of nerfl ets so as to form a decomposed representation of the scene, where each object in stance is represented by a group of nerflets. During experiments with indoor and outdoor environments, we find that nerflets: (1) fit and approximate the scene more efficiently than traditional global NeRFs, (2) allow the extraction of pano ptic and photometric renderings from arbitrary views, and (3) enable tasks rare for NeRFs, such as 3D panoptic segmentation and interactive editing.
****
CLIP Is Also an Efficient Segmenter: A Text-Driven Approach for Weakly Supervise d Semantic Segmentation
Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, X iaofei He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15305-15314
Weakly supervised semantic segmentation (WSSS) with image-level labels is a chal lenging task. Mainstream approaches follow a multi-stage framework and suffer fr om high training costs. In this paper, we explore the potential of Contrastive L anguage-Image Pre-training models (CLIP) to localize different categories with o nly image-level labels and without further training. To efficiently generate hig h-quality segmentation masks from CLIP, we propose a novel WSSS framework called CLIP-ES. Our framework improves all three stages of WSSS with special designs f or CLIP: 1) We introduce the softmax function into GradCAM and exploit the zero- shot ability of CLIP to suppress the confusion caused by non-target classes and backgrounds. Meanwhile, to take full advantage of CLIP, we re-explore text input s under the WSSS setting and customize two text-driven strategies: sharpness-bas ed prompt selection and synonym fusion. 2) To simplify the stage of CAM refineme nt, we propose a real-time class-aware attention-based affinity (CAA) module bas ed on the inherent multi-head self-attention (MHSA) in CLIP-ViTs. 3) When traini ng the final segmentation model with the masks generated by CLIP, we introduced a confidence-guided loss (CGL) focus on confident regions. Our CLIP-ES achieves SOTA performance on Pascal VOC 2012 and MS COCO 2014 while only taking 10% time of previous methods for the pseudo mask generation. Code is available at https:/ /github.com/linyq2117/CLIP-ES.
****
Slimmable Dataset Condensation
Songhua Liu, Jingwen Ye, Runpeng Yu, Xinchao Wang; Proceedings of the IEEE/CVF C onference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3759-3768
Dataset distillation, also known as dataset condensation, aims to compress a lar ge dataset into a compact synthetic one. Existing methods perform dataset conden sation by assuming a fixed storage or transmission budget. When the budget chang es, however, they have to repeat the synthesizing process with access to origina l datasets, which is highly cumbersome if not infeasible at all. In this paper, we explore the problem of slimmable dataset condensation, to extract a smaller s ynthetic dataset given only previous condensation results. We first study the li mitations of existing dataset condensation algorithms on such a successive compr ession setting and identify two key factors: (1) the inconsistency of neural net works over different compression times and (2) the underdetermined solution spac e for synthetic data. Accordingly, we propose a novel training objective for sli

mmable dataset condensation to explicitly account for both factors. Moreover, sy nthetic datasets in our method adopt an significance-aware parameterization. The oretical derivation indicates that an upper-bounded error can be achieved by dis carding the minor components without training. Alternatively, if training is all owed, this strategy can serve as a strong initialization that enables a fast con vergence. Extensive comparisons and ablations demonstrate the superiority of the proposed solution over existing methods on multiple benchmarks.
****
Spatially Adaptive Self-Supervised Learning for Real-World Image Denoising
Junyi Li, Zhilu Zhang, Xiaoyu Liu, Chaoyu Feng, Xiaotao Wang, Lei Lei, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2023, pp. 9914-9924
Significant progress has been made in self-supervised image denoising (SSID) in the recent few years. However, most methods focus on dealing with spatially inde pendent noise, and they have little practicality on real-world sRGB images with spatially correlated noise. Although pixel-shuffle downsampling has been suggest ed for breaking the noise correlation, it breaks the original information of ima ges, which limits the denoising performance. In this paper, we propose a novel p erspective to solve this problem, i.e., seeking for spatially adaptive supervisi on for real-world sRGB image denoising. Specifically, we take into account the r espective characteristics of flat and textured regions in noisy images, and cons truct supervisions for them separately. For flat areas, the supervision can be s afely derived from non-adjacent pixels, which are much far from the current pixe l for excluding the influence of the noise-correlated ones. And we extend the bl ind-spot network to a blind-neighborhood network (BNN) for providing supervision on flat areas. For textured regions, the supervision has to be closely related to the content of adjacent pixels. And we present a locally aware network (LAN) to meet the requirement, while LAN itself is selectively supervised with the out put of BNN. Combining these two supervisions, a denoising network (e.g., U-Net) can be well-trained. Extensive experiments show that our method performs favorab ly against state-of-the-art SSID methods on real-world sRGB photographs. The cod e is available at https://github.com/nagejacob/SpatiallyAdaptiveSSID.
****
Data-Free Knowledge Distillation via Feature Exchange and Activation Region Cons traint
Shikang Yu, Jiachen Chen, Hu Han, Shuqiang Jiang; Proceedings of the IEEE/CVF Co nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24266-2427 5
Despite the tremendous progress on data-free knowledge distillation (DFKD) based on synthetic data generation, there are still limitations in diverse and effici ent data synthesis. It is naive to expect that a simple combination of generativ e network-based data synthesis and data augmentation will solve these issues. Th erefore, this paper proposes a novel data-free knowledge distillation method (Sp aceshipNet) based on channel-wise feature exchange (CFE) and multi-scale spatial activation region consistency (mSARC) constraint. Specifically, CFE allows our generative network to better sample from the feature space and efficiently synth esize diverse images for learning the student network. However, using CFE alone can severely amplify the unwanted noises in the synthesized images, which may re sult in failure to improve distillation learning and even have negative effects. Therefore, we propose mSARC to assure the student network can imitate not only the logit output but also the spatial activation region of the teacher network i n order to alleviate the influence of unwanted noises in diverse synthetic image s on distillation learning. Extensive experiments on CIFAR-10, CIFAR-100, Tiny-I mageNet, Imagenette, and ImageNet100 show that our method can work well with dif ferent backbone networks, and outperform the state-of-the-art DFKD methods. Code will be available at: https://github.com/skgyu/SpaceshipNet.
****
CLIP-Sculptor: Zero-Shot Generation of High-Fidelity and Diverse Shapes From Nat ural Language
Aditya Sanghi, Rao Fu, Vivian Liu, Karl D.D. Willis, Hooman Shayani, Amir H. Kha

sahmadi, Srinath Sridhar, Daniel Ritchie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18339-18348
Recent works have demonstrated that natural language can be used to generate and edit 3D shapes. However, these methods generate shapes with limited fidelity and diversity. We introduce CLIP-Sculptor, a method to address these constraints by producing high-fidelity and diverse 3D shapes without the need for (text, shape) pairs during training. CLIP-Sculptor achieves this in a multi-resolution approach that first generates in a low-dimensional latent space and then upscales to a higher resolution for improved shape fidelity. For improved shape diversity, we use a discrete latent space which is modeled using a transformer conditioned on CLIP's image-text embedding space. We also present a novel variant of classifier-free guidance, which improves the accuracy-diversity trade-off. Finally, we perform extensive experiments demonstrating that CLIP-Sculptor outperforms state-of-the-art baselines.
****
Mask-Free Video Instance Segmentation
Lei Ke, Martin Danelljan, Henghui Ding, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22857-22866
The recent advancement in Video Instance Segmentation (VIS) has largely been driven by the use of deeper and increasingly data-hungry transformer-based models. However, video masks are tedious and expensive to annotate, limiting the scale and diversity of existing VIS datasets. In this work, we aim to remove the mask-annotation requirement. We propose MaskFreeVIS, achieving highly competitive VIS performance, while only using bounding box annotations for the object state. We leverage the rich temporal mask consistency constraints in videos by introducing the Temporal KNN-patch Loss (TK-Loss), providing strong mask supervision without any labels. Our TK-Loss finds one-to-many matches across frames, through an efficient patch-matching step followed by a K-nearest neighbor selection. A consistency loss is then enforced on the found matches. Our mask-free objective is simple to implement, has no trainable parameters, is computationally efficient, yet outperforms baselines employing, e.g., state-of-the-art optical flow to enforce temporal mask consistency. We validate MaskFreeVIS on the YouTube-VIS 2019/2021, OVIS and BDD100K MOTS benchmarks. The results clearly demonstrate the efficacy of our method by drastically narrowing the gap between fully and weakly-supervised VIS performance. Our code and trained models are available at http://vis.xyz/pub/maskfreevis.
****
Continual Detection Transformer for Incremental Object Detection
Yaoyao Liu, Bernt Schiele, Andrea Vedaldi, Christian Rupprecht; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23799-23808
Incremental object detection (IOD) aims to train an object detector in phases, each with annotations for new object categories. As other incremental settings, IOD is subject to catastrophic forgetting, which is often addressed by techniques such as knowledge distillation (KD) and exemplar replay (ER). However, KD and ER do not work well if applied directly to state-of-the-art transformer-based object detectors such as Deformable DETR and UP-DETR. In this paper, we solve these issues by proposing a ContinuaL DEtection TRansformer (CL-DETR), a new method for transformer-based IOD which enables effective usage of KD and ER in this context. First, we introduce a Detector Knowledge Distillation (DKD) loss, focusing on the most informative and reliable predictions from old versions of the model, ignoring redundant background predictions, and ensuring compatibility with the available ground-truth labels. We also improve ER by proposing a calibration strategy to preserve the label distribution of the training set, therefore better matching training and testing statistics. We conduct extensive experiments on COCO 2017 and demonstrate that CL-DETR achieves state-of-the-art results in the IOD setting.
****
Two-Stream Networks for Weakly-Supervised Temporal Action Localization With Sema

ntic-Aware Mechanisms
Yu Wang, Yadong Li, Hongbin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18878-18887
Weakly-supervised temporal action localization aims to detect action boundaries in untrimmed videos with only video-level annotations. Most existing schemes detect temporal regions that are most responsive to video-level classification, but they overlook the semantic consistency between frames. In this paper, we hypothesize that snippets with similar representations should be considered as the same action class despite the absence of supervision signals on each snippet. To this end, we devise a learnable dictionary where entries are the class centroids of the corresponding action categories. The representations of snippets identified as the same action category are induced to be close to the same class centroid, which guides the network to perceive the semantics of frames and avoid unreasonable localization. Besides, we propose a two-stream framework that integrates the attention mechanism and the multiple-instance learning strategy to extract fine-grained clues and salient features respectively. Their complementarity enables the model to refine temporal boundaries. Finally, the developed model is validated on the publicly available THUMOS-14 and ActivityNet-1.3 datasets, where substantial experiments and analyses demonstrate that our model achieves remarkable advances over existing methods.
****
HyperMatch: Noise-Tolerant Semi-Supervised Learning via Relaxed Contrastive Constraint
Beitong Zhou, Jing Lu, Kerui Liu, Yunlu Xu, Zhanzhan Cheng, Yi Niu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24017-24026
Recent developments of the application of Contrastive Learning in Semi-Supervised Learning (SSL) have demonstrated significant advancements, as a result of its exceptional ability to learn class-aware cluster representations and the full exploitation of massive unlabeled data. However, mismatched instance pairs caused by inaccurate pseudo labels would assign an unlabeled instance to the incorrect class in feature space, hence exacerbating SSL's renowned confirmation bias. To address this issue, we introduced a novel SSL approach, HyperMatch, which is a plug-in to several SSL designs enabling noise-tolerant utilization of unlabeled data. In particular, confidence predictions are combined with semantic similarities to generate a more objective class distribution, followed by a Gaussian Mixture Model to divide pseudo labels into a 'confident' and a 'less confident' subset. Then, we introduce Relaxed Contrastive Loss by assigning the 'less-confident' samples to a hyper-class, i.e. the union of top-K nearest classes, which effectively regularizes the interference of incorrect pseudo labels and even increases the probability of pulling a 'less confident' sample close to its true class. Experiments and in-depth studies demonstrate that HyperMatch delivers remarkable state-of-the-art performance, outperforming FixMatch on CIFAR100 with 400 and 2500 labeled samples by 11.86% and 4.88%, respectively.
****
From Images to Textual Prompts: Zero-Shot Visual Question Answering With Frozen Large Language Models
Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, Steven Hoi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10867-10877
Large language models (LLMs) have demonstrated excellent zero-shot generalization to new language tasks. However, effective utilization of LLMs for zero-shot visual question-answering (VQA) remains challenging, primarily due to the modality disconnection and task disconnection between LLM and VQA task. End-to-end training on vision and language data may bridge the disconnections, but is inflexible and computationally expensive. To address this issue, we propose Img2Prompt, a plug-and-play module that provides the prompts that can bridge the aforementioned modality and task disconnections, so that LLMs can perform zero-shot VQA tasks without end-to-end training. In order to provide such prompts, we further employ LLM-agnostic models to provide prompts that can describe image content and sel

f-constructed question-answer pairs, which can effectively guide LLM to perform zero-shot VQA tasks. Img2Prompt offers the following benefits: 1) It can flexibly work with various LLMs to perform VQA. 2) Without the needing of end-to-end training, it significantly reduces the cost of deploying LLM for zero-shot VQA tasks. 3) It achieves comparable or better performance than methods relying on end-to-end training. For example, we outperform Flamingo by 5.6% on VQAv2. On the challenging A-OKVQA dataset, our method even outperforms few-shot methods by as much as 20%.
****
LEGO-Net: Learning Regular Rearrangements of Objects in Rooms
Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, Leonidas Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19037-19047
Humans universally dislike the task of cleaning up a messy room. If machines were to help us with this task, they must understand human criteria for regular arrangements, such as several types of symmetry, co-linearity or co-circularity, spacing uniformity in linear or circular patterns, and further inter-object relationships that relate to style and functionality. Previous approaches for this task relied on human input to explicitly specify goal state, or synthesized scenes from scratch--but such methods do not address the rearrangement of existing messy scenes without providing a goal state. In this paper, we present LEGO-Net, a data-driven transformer-based iterative method for LEarning reGular rearrangement of Objects in messy rooms. LEGO-Net is partly inspired by diffusion models--it starts with an initial messy state and iteratively "de-noises" the position and orientation of objects to a regular state while reducing distance traveled. Given randomly perturbed object positions and orientations in an existing dataset of professionally-arranged scenes, our method is trained to recover a regular re-arrangement. Results demonstrate that our method is able to reliably rearrange room scenes and outperform other methods. We additionally propose a metric for evaluating regularity in room arrangements using number-theoretic machinery.
****
FastInst: A Simple Query-Based Model for Real-Time Instance Segmentation
Junjie He, Pengyu Li, Yifeng Geng, Xuansong Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23663-23672
Recent attention in instance segmentation has focused on query-based models. Despite being non-maximum suppression (NMS)-free and end-to-end, the superiority of these models on high-accuracy real-time benchmarks has not been well demonstrated. In this paper, we show the strong potential of query-based models on efficient instance segmentation algorithm designs. We present FastInst, a simple, effective query-based framework for real-time instance segmentation. FastInst can execute at a real-time speed (i.e., 32.5 FPS) while yielding an AP of more than 40 (i.e., 40.5 AP) on COCO test-dev without bells and whistles. Specifically, FastInst follows the meta-architecture of recently introduced Mask2Former. Its key designs include instance activation-guided queries, dual-path update strategy, and ground truth mask-guided learning, which enable us to use lighter pixel decoders, fewer Transformer decoder layers, while achieving better performance. The experiments show that FastInst outperforms most state-of-the-art real-time counterparts, including strong fully convolutional baselines, in both speed and accuracy. Code can be found at https://github.com/junjiehe96/FastInst.
****
Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking
Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, Kris Kitani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9686-9696
Kalman filter (KF) based methods for multi-object tracking (MOT) make an assumption that objects move linearly. While this assumption is acceptable for very short periods of occlusion, linear estimates of motion for prolonged time can be highly inaccurate. Moreover, when there is no measurement available to update Kalman filter parameters, the standard convention is to trust the priori state estimations for posteriori update. This leads to the accumulation of errors during a

period of occlusion. The error causes significant motion direction variance in practice. In this work, we show that a basic Kalman filter can still obtain state-of-the-art tracking performance if proper care is taken to fix the noise accumulated during occlusion. Instead of relying only on the linear state estimate (i.e., estimation-centric approach), we use object observations (i.e., the measurements by object detector) to compute a virtual trajectory over the occlusion period to fix the error accumulation of filter parameters. This allows more time steps to correct errors accumulated during occlusion. We name our method Observation-Centric SORT (OC-SORT). It remains Simple, Online, and Real-Time but improves robustness during occlusion and non-linear motion. Given off-the-shelf detections as input, OC-SORT runs at 700+ FPS on a single CPU. It achieves state-of-the-art on multiple datasets, including MOT17, MOT20, KITTI, head tracking, and especially DanceTrack where the object motion is highly non-linear. The code and models are available at https://github.com/noahcao/OC_SORT.
****
Multi-View Azimuth Stereo via Tangent Space Consistency
Xu Cao, Hiroaki Santo, Fumio Okura, Yasuyuki Matsushita; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 825-834
We present a method for 3D reconstruction only using calibrated multi-view surface azimuth maps. Our method, multi-view azimuth stereo, is effective for textureless or specular surfaces, which are difficult for conventional multi-view stereo methods. We introduce the concept of tangent space consistency: Multi-view azimuth observations of a surface point should be lifted to the same tangent space. Leveraging this consistency, we recover the shape by optimizing a neural implicit surface representation. Our method harnesses the robust azimuth estimation capabilities of photometric stereo methods or polarization imaging while bypassing potentially complex zenith angle estimation. Experiments using azimuth maps from various sources validate the accurate shape recovery with our method, even without zenith angles.
****
VectorFusion: Text-to-SVG by Abstracting Pixel-Based Diffusion Models
Ajay Jain, Amber Xie, Pieter Abbeel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1911-1920
Diffusion models have shown impressive results in text-to-image synthesis. Using massive datasets of captioned images, diffusion models learn to generate raster images of highly diverse objects and scenes. However, designers frequently use vector representations of images like Scalable Vector Graphics (SVGs) for digital icons, graphics and stickers. Vector graphics can be scaled to any size, and are compact. In this work, we show that a text-conditioned diffusion model trained on pixel representations of images can be used to generate SVG-exportable vector graphics. We do so without access to large datasets of captioned SVGs. Instead, inspired by recent work on text-to-3D synthesis, we vectorize a text-to-image diffusion sample and fine-tune with a Score Distillation Sampling loss. By optimizing a differentiable vector graphics rasterizer, our method distills abstract semantic knowledge out of a pretrained diffusion model. By constraining the vector representation, we can also generate coherent pixel art and sketches. Our approach, VectorFusion, produces more coherent graphics than prior works that optimize CLIP, a contrastive image-text model.
****
The Dialog Must Go On: Improving Visual Dialog via Generative Self-Training
Gi-Cheon Kang, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, Byoung-Tak Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6746-6756
Visual dialog (VisDial) is a task of answering a sequence of questions grounded in an image, using the dialog history as context. Prior work has trained the dialog agents solely on VisDial data via supervised learning or leveraged pre-training on related vision-and-language datasets. This paper presents a semi-supervised learning approach for visually-grounded dialog, called Generative Self-Training (GST), to leverage unlabeled images on the Web. Specifically, GST first retri

eves in-domain images through out-of-distribution detection and generates synthe
tic dialogs regarding the images via multimodal conditional text generation. GST then trains a dialog agent on the synthetic and the original VisDial data. As a result, GST scales the amount of training data up to an order of magnitude that of VisDial (1.2M to 12.9M QA data). For robust training of the synthetic dialogs, we also propose perplexity-based data selection and multimodal consistency regularization. Evaluation on VisDial v1.0 and v0.9 datasets shows that GST achieves new state-of-the-art results on both datasets. We further observe the robustness of GST against both visual and textual adversarial attacks. Finally, GST yields strong performance gains in the low-data regime. Code is available at https://github.com/gicheonkang/gst-visdial.
****
Binarizing Sparse Convolutional Networks for Efficient Point Cloud Analysis
Xiuwei Xu, Ziwei Wang, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5313-5322
In this paper, we propose binary sparse convolutional networks called BSC-Net for efficient point cloud analysis. We empirically observe that sparse convolution operation causes larger quantization errors than standard convolution. However, conventional network quantization methods directly binarize the weights and activations in sparse convolution, resulting in performance drop due to the significant quantization loss. On the contrary, we search the optimal subset of convolution operation that activates the sparse convolution at various locations for quantization error alleviation, and the performance gap between real-valued and binary sparse convolutional networks is closed without complexity overhead. Specifically, we first present the shifted sparse convolution that fuses the information in the receptive field for the active sites that match the pre-defined positions. Then we employ the differentiable search strategies to discover the optimal opsitions for active site matching in the shifted sparse convolution, and the quantization errors are significantly alleviated for efficient point cloud analysis. For fair evaluation of the proposed method, we empirically select the recently advances that are beneficial for sparse convolution network binarization to construct a strong baseline. The experimental results on ScanNet and NYU Depth v2 show that our BSC-Net achieves significant improvement upon our srtong baseline and outperforms the state-of-the-art network binarization methods by a remarkable margin without additional computation overhead for binarizing sparse convolutional networks.
****
Transformer-Based Learned Optimization
Erik Gärtner, Luke Metz, Mykhaylo Andriluka, C. Daniel Freeman, Cristian Sminchisescu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11970-11979
We propose a new approach to learned optimization where we represent the computation of an optimizer's update step using a neural network. The parameters of the optimizer are then learned by training on a set of optimization tasks with the objective to perform minimization efficiently. Our innovation is a new neural network architecture, Optimus, for the learned optimizer inspired by the classic BFGS algorithm. As in BFGS, we estimate a preconditioning matrix as a sum of rank-one updates but use a Transformer-based neural network to predict these updates jointly with the step length and direction. In contrast to several recent learned optimization-based approaches, our formulation allows for conditioning across the dimensions of the parameter space of the target problem while remaining applicable to optimization tasks of variable dimensionality without retraining. We demonstrate the advantages of our approach on a benchmark composed of objective functions traditionally used for the evaluation of optimization algorithms, as well as on the real world-task of physics-based visual reconstruction of articulated 3d human motion.
****
Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models
Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, Tom Goldstein; P

roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6048-6058
Cutting-edge diffusion models produce images with high quality and customizability, enabling them to be used for commercial art and graphic design purposes. But do diffusion models create unique works of art, or are they replicating content directly from their training sets? In this work, we study image retrieval frameworks that enable us to compare generated images with training samples and detect when content has been replicated. Applying our frameworks to diffusion models trained on multiple datasets including Oxford flowers, Celeb-A, ImageNet, and LAION, we discuss how factors such as training set size impact rates of content replication. We also identify cases where diffusion models, including the popular Stable Diffusion model, blatantly copy from their training data.
****
Neuralizer: General Neuroimage Analysis Without Re-Training
Steffen Czolbe, Adrian V. Dalca; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6217-6230
Neuroimage processing tasks like segmentation, reconstruction, and registration are central to the study of neuroscience. Robust deep learning strategies and architectures used to solve these tasks are often similar. Yet, when presented with a new task or a dataset with different visual characteristics, practitioners most often need to train a new model, or fine-tune an existing one. This is a time-consuming process that poses a substantial barrier for the thousands of neuroscientists and clinical researchers who often lack the resources or machine-learning expertise to train deep learning models. In practice, this leads to a lack of adoption of deep learning, and neuroscience tools being dominated by classical frameworks. We introduce Neuralizer, a single model that generalizes to previously unseen neuroimaging tasks and modalities without the need for re-training or fine-tuning. Tasks do not have to be known a priori, and generalization happens in a single forward pass during inference. The model can solve processing tasks across multiple image modalities, acquisition methods, and datasets, and generalize to tasks and modalities it has not been trained on. Our experiments on coronal slices show that when few annotated subjects are available, our multi-task network outperforms task-specific baselines without training on the task.
****
Quantum-Inspired Spectral-Spatial Pyramid Network for Hyperspectral Image Classification
Jie Zhang, Yongshan Zhang, Yicong Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9925-9934
Hyperspectral image (HSI) classification aims at assigning a unique label for every pixel to identify categories of different land covers. Existing deep learning models for HSIs are usually performed in a traditional learning paradigm. Being emerging machines, quantum computers are limited in the noisy intermediate-scale quantum (NISQ) era. The quantum theory offers a new paradigm for designing deep learning models. Motivated by the quantum circuit (QC) model, we propose a quantum-inspired spectral-spatial network (QSSN) for HSI feature extraction. The proposed QSSN consists of a phase-prediction module (PPM) and a measurement-like fusion module (MFM) inspired from quantum theory to dynamically fuse spectral and spatial information. Specifically, QSSN uses a quantum representation to represent an HSI cuboid and extracts joint spectral-spatial features using MFM. An HSI cuboid and its phases predicted by PPM are used in the quantum representation. Using QSSN as the building block, we propose an end-to-end quantum-inspired spectral-spatial pyramid network (QSSPN) for HSI feature extraction and classification. In this pyramid framework, QSSPN progressively learns feature representations by cascading QSSN blocks and performs classification with a softmax classifier. It is the first attempt to introduce quantum theory in HSI processing model design. Substantial experiments are conducted on three HSI datasets to verify the superiority of the proposed QSSPN framework over the state-of-the-art methods.
****
Towards Benchmarking and Assessing Visual Naturalness of Physical World Adversarial Attacks

Simin Li, Shuning Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, Xianglong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12324-12333
Physical world adversarial attack is a highly practical and threatening attack, which fools real world deep learning systems by generating conspicuous and maliciously crafted real world artifacts. In physical world attacks, evaluating naturalness is highly emphasized since human can easily detect and remove unnatural attacks. However, current studies evaluate naturalness in a case-by-case fashion, which suffers from errors, bias and inconsistencies. In this paper, we take the first step to benchmark and assess visual naturalness of physical world attacks, taking autonomous driving scenario as the first attempt. First, to benchmark attack naturalness, we contribute the first Physical Attack Naturalness (PAN) dataset with human rating and gaze. PAN verifies several insights for the first time: naturalness is (disparately) affected by contextual features (i.e., environmental and semantic variations) and correlates with behavioral feature (i.e., gaze signal). Second, to automatically assess attack naturalness that aligns with human ratings, we further introduce Dual Prior Alignment (DPA) network, which aims to embed human knowledge into model reasoning process. Specifically, DPA imitates human reasoning in naturalness assessment by rating prior alignment and mimics human gaze behavior by attentive prior alignment. We hope our work fosters researches to improve and automatically assess naturalness of physical world attacks. Our code and exemplar data can be found at https://github.com/zhangsn-19/PAN.
****
Visual Prompt Multi-Modal Tracking
Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, Huchuan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9516-9526
Visible-modal object tracking gives rise to a series of downstream multi-modal tracking tributaries. To inherit the powerful representations of the foundation model, a natural modus operandi for multi-modal tracking is full fine-tuning on the RGB-based parameters. Albeit effective, this manner is not optimal due to the scarcity of downstream data and poor transferability, etc. In this paper, inspired by the recent success of the prompt learning in language models, we develop Visual Prompt multi-modal Tracking (ViPT), which learns the modal-relevant prompts to adapt the frozen pre-trained foundation model to various downstream multimodal tracking tasks. ViPT finds a better way to stimulate the knowledge of the RGB-based model that is pre-trained at scale, meanwhile only introducing a few trainable parameters (less than 1% of model parameters). ViPT outperforms the full fine-tuning paradigm on multiple downstream tracking tasks including RGB+Depth, RGB+Thermal, and RGB+Event tracking. Extensive experiments show the potential of visual prompt learning for multi-modal tracking, and ViPT can achieve state-of-the-art performance while satisfying parameter efficiency. Code and models are available at https://github.com/jiawen-zhu/ViPT.
****
Self-Supervised Representation Learning for CAD
Benjamin T. Jones, Michael Hu, Milin Kodnongbua, Vladimir G. Kim, Adriana Schulz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21327-21336
Virtually every object in the modern world was created, modified, analyzed and optimized using computer aided design (CAD) tools. An active CAD research area is the use of data-driven machine learning methods to learn from the massive repositories of geometric and program representations. However, the lack of labeled data in CAD's native format, i.e., the parametric boundary representation (B-Rep), poses an obstacle at present difficult to overcome. Several datasets of mechanical parts in B-Rep format have recently been released for machine learning research. However, large-scale databases are mostly unlabeled, and labeled datasets are small. Additionally, task-specific label sets are rare and costly to annotate. This work proposes to leverage unlabeled CAD geometry on supervised learning tasks. We learn a novel, hybrid implicit/explicit surface representation for B-Rep geometry. Further, we show that this pre-training both significantly improves

few-shot learning performance and achieves state-of-the-art performance on several current B-Rep benchmarks.
****
DETRs With Hybrid Matching
Ding Jia, Yuhui Yuan, Haodi He, Xiaopei Wu, Haojun Yu, Weihong Lin, Lei Sun, Chao Zhang, Han Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19702-19712
One-to-one set matching is a key design for DETR to establish its end-to-end capability, so that object detection does not require a hand-crafted NMS (non-maximum suppression) to remove duplicate detections. This end-to-end signature is important for the versatility of DETR, and it has been generalized to broader vision tasks. However, we note that there are few queries assigned as positive samples and the one-to-one set matching significantly reduces the training efficacy of positive samples. We propose a simple yet effective method based on a hybrid matching scheme that combines the original one-to-one matching branch with an auxiliary one-to-many matching branch during training. Our hybrid strategy has been shown to significantly improve accuracy. In inference, only the original one-to-one match branch is used, thus maintaining the end-to-end merit and the same inference efficiency of DETR. The method is named H-DETR, and it shows that a wide range of representative DETR methods can be consistently improved across a wide range of visual tasks, including Deformable-DETR, PETRv2, PETR, and TransTrack, among others.
****
Dealing With Cross-Task Class Discrimination in Online Continual Learning
Yiduo Guo, Bing Liu, Dongyan Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11878-11887
Existing continual learning (CL) research regards catastrophic forgetting (CF) as almost the only challenge. This paper argues for another challenge in class-incremental learning (CIL), which we call cross-task class discrimination (CTCD), i.e., how to establish decision boundaries between the classes of the new task and old tasks with no (or limited) access to the old task data. CTCD is implicitly and partially dealt with by replay-based methods. A replay method saves a small amount of data (replay data) from previous tasks. When a batch of current task data arrives, the system jointly trains the new data and some sampled replay data. The replay data enables the system to partially learn the decision boundaries between the new classes and the old classes as the amount of the saved data is small. However, this paper argues that the replay approach also has a dynamic training bias issue which reduces the effectiveness of the replay data in solving the CTCD problem. A novel optimization objective with a gradient-based adaptive method is proposed to dynamically deal with the problem in the online CL process. Experimental results show that the new method achieves much better results in online CL.
****
Angelic Patches for Improving Third-Party Object Detector Performance
Wenwen Si, Shuo Li, Sangdon Park, Insup Lee, Osbert Bastani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24638-24647
Deep learning models have shown extreme vulnerability to simple perturbations and spatial transformations. In this work, we explore whether we can adopt the characteristics of adversarial attack methods to help improve perturbation robustness for object detection. We study a class of realistic object detection settings wherein the target objects have control over their appearance. To this end, we propose a reversed Fast Gradient Sign Method (FGSM) to obtain these angelic patches that significantly increase the detection probability, even without pre-knowledge of the perturbations. In detail, we apply the patch to each object instance simultaneously, strengthen not only classification but also bounding box accuracy. Experiments demonstrate the efficacy of the partial-covering patch in solving the complex bounding box problem. More importantly, the performance is also transferable to different detection models even under severe affine transformations and deformable shapes. To our knowledge, we are the first (object detection)

patch that achieves both cross-model and multiple-patch efficacy. We observed av erage accuracy improvements of 30% in the real-world experiments, which brings l arge social value. Our code is available at: https://github.com/averysi224/angel ic_patches.
****
UniDexGrasp: Universal Robotic Dexterous Grasping via Learning Diverse Proposal Generation and Goal-Conditioned Policy
Yinzhen Xu, Weikang Wan, Jialiang Zhang, Haoran Liu, Zikang Shan, Hao Shen, Ruic heng Wang, Haoran Geng, Yijia Weng, Jiayi Chen, Tengyu Liu, Li Yi, He Wang; Proc eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C VPR), 2023, pp. 4737-4746
In this work, we tackle the problem of learning universal robotic dexterous gras ping from a point cloud observation under a table-top setting. The goal is to gr asp and lift up objects in high-quality and diverse ways and generalize across h undreds of categories and even the unseen. Inspired by successful pipelines used in parallel gripper grasping, we split the task into two stages: 1) grasp propo sal (pose) generation and 2) goal-conditioned grasp execution. For the first sta ge, we propose a novel probabilistic model of grasp pose conditioned on the poin t cloud observation that factorizes rotation from translation and articulation. Trained on our synthesized large-scale dexterous grasp dataset, this model enabl es us to sample diverse and high-quality dexterous grasp poses for the object po int cloud. For the second stage, we propose to replace the motion planning used in parallel gripper grasping with a goal-conditioned grasp policy, due to the co mplexity involved in dexterous grasping execution. Note that it is very challeng ing to learn this highly generalizable grasp policy that only takes realistic in puts without oracle states. We thus propose several important innovations, inclu ding state canonicalization, object curriculum, and teacher-student distillation . Integrating the two stages, our final pipeline becomes the first to achieve un iversal generalization for dexterous grasping, demonstrating an average success rate of more than 60% on thousands of object instances, which significantly outp erforms all baselines, meanwhile showing only a minimal generalization gap.
****
A Rotation-Translation-Decoupled Solution for Robust and Efficient Visual-Inerti al Initialization
Yijia He, Bo Xu, Zhanpeng Ouyang, Hongdong Li; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 739-748
We propose a novel visual-inertial odometry (VIO) initialization method, which d ecouples rotation and translation estimation, and achieves higher efficiency and better robustness. Existing loosely-coupled VIO-initialization methods suffer f rom poor stability of visual structure-from-motion (SfM), whereas those tightly-coupled methods often ignore the gyroscope bias in the closed-form solution, res ulting in limited accuracy. Moreover, the aforementioned two classes of methods are computationally expensive, because 3D point clouds need to be reconstructed simultaneously. In contrast, our new method fully combines inertial and visual m easurements for both rotational and translational initialization. First, a rotat ion-only solution is designed for gyroscope bias estimation, which tightly coupl es the gyroscope and camera observations. Second, the initial velocity and gravi ty vector are solved with linear translation constraints in a globally optimal f ashion and without reconstructing 3D point clouds. Extensive experiments have de monstrated that our method is 8 72 times faster (w.r.t. a 10-frame set) than the state-of-the-art methods, and also presents significantly higher robustness and accuracy. The source code is available at https://github.com/boxuLibrary/drt-vi o-init.
****
GIVL: Improving Geographical Inclusivity of Vision-Language Models With Pre-Trai ning Methods
Da Yin, Feng Gao, Govind Thattai, Michael Johnston, Kai-Wei Chang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 3, pp. 10951-10961
A key goal for the advancement of AI is to develop technologies that serve the n

eeds not just of one group but of all communities regardless of their geographical region. In fact, a significant proportion of knowledge is locally shared by people from certain regions but may not apply equally in other regions because of cultural differences. If a model is unaware of regional characteristics, it may lead to performance disparity across regions and result in bias against underrepresented groups. We propose GIVL, a Geographically Inclusive Vision-and-Language Pre-trained model. There are two attributes of geo-diverse visual concepts which can help to learn geo-diverse knowledge: 1) concepts under similar categories have unique knowledge and visual characteristics, 2) concepts with similar visual features may fall in completely different categories. Motivated by the attributes, we design new pre-training objectives Image-Knowledge Matching (IKM) and Image Edit Checking (IEC) to pre-train GIVL. Compared with similar-size models pre-trained with similar scale of data, GIVL achieves state-of-the-art (SOTA) and more balanced performance on geo-diverse V&L tasks. Code and data are released at https://github.com/WadeYin9712/GIVL.
****
Bi3D: Bi-Domain Active Learning for Cross-Domain 3D Object Detection
Jiakang Yuan, Bo Zhang, Xiangchao Yan, Tao Chen, Botian Shi, Yikang Li, Yu Qiao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15599-15608
Unsupervised Domain Adaptation (UDA) technique has been explored in 3D cross-domain tasks recently. Though preliminary progress has been made, the performance gap between the UDA-based 3D model and the supervised one trained with fully annotated target domain is still large. This motivates us to consider selecting partial-yet-important target data and labeling them at a minimum cost, to achieve a good trade-off between high performance and low annotation cost. To this end, we propose a Bi-domain active learning approach, namely Bi3D, to solve the cross-domain 3D object detection task. The Bi3D first develops a domainness-aware source sampling strategy, which identifies target-domain-like samples from the source domain to avoid the model being interfered by irrelevant source data. Then a diversity-based target sampling strategy is developed, which selects the most informative subset of target domain to improve the model adaptability to the target domain using as little annotation budget as possible. Experiments are conducted on typical cross-domain adaptation scenarios including cross-LiDAR-beam, cross-country, and cross-sensor, where Bi3D achieves a promising target-domain detection accuracy (89.63% on KITTI) compared with UDA-based work (84.29%), even surpassing the detector trained on the full set of the labeled target domain (88.98%).
****
Towards Fast Adaptation of Pretrained Contrastive Models for Multi-Channel Video-Language Retrieval
Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, Shih-Fu Chang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14846-14855
Multi-channel video-language retrieval require models to understand information from different channels (e.g. video+question, video+speech) to correctly link a video with a textual response or query. Fortunately, contrastive multimodal models are shown to be highly effective at aligning entities in images/videos and text, e.g., CLIP; text contrastive models are extensively studied recently for their strong ability of producing discriminative sentence embeddings, e.g., SimCSE. However, there is not a clear way to quickly adapt these two lines to multi-channel video-language retrieval with limited data and resources. In this paper, we identify a principled model design space with two axes: how to represent videos and how to fuse video and text information. Based on categorization of recent methods, we investigate the options of representing videos using continuous feature vectors or discrete text tokens; for the fusion method, we explore the use of a multimodal transformer or a pretrained contrastive text model. We extensively evaluate the four combinations on five video-language datasets. We surprisingly find that discrete text tokens coupled with a pretrained contrastive text model yields the best performance, which can even outperform state-of-the-art on the iVQA and How2QA datasets without additional training on millions of video-text d

ata. Further analysis shows that this is because representing videos as text tokens captures the key visual information and text tokens are naturally aligned with text models that are strong retrievers after the contrastive pretraining process. All the empirical analysis establishes a solid foundation for future research on affordable and upgradable multimodal intelligence. The code will be released at https://github.com/XudongLinthu/upgradable-multimodal-intelligence to facilitate future research.
****
Mask-Free OVIS: Open-Vocabulary Instance Segmentation Without Manual Mask Annotations
Vibashan VS, Ning Yu, Chen Xing, Can Qin, Mingfei Gao, Juan Carlos Niebles, Vishal M. Patel, Ran Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23539-23549
Existing instance segmentation models learn task-specific information using manual mask annotations from base (training) categories. These mask annotations require tremendous human effort, limiting the scalability to annotate novel (new) categories. To alleviate this problem, Open-Vocabulary (OV) methods leverage large-scale image-caption pairs and vision-language models to learn novel categories. In summary, an OV method learns task-specific information using strong supervision from base annotations and novel category information using weak supervision from image-captions pairs. This difference between strong and weak supervision leads to overfitting on base categories, resulting in poor generalization towards novel categories. In this work, we overcome this issue by learning both base and novel categories from pseudo-mask annotations generated by the vision-language model in a weakly supervised manner using our proposed Mask-free OVIS pipeline. Our method automatically generates pseudo-mask annotations by leveraging the localization ability of a pre-trained vision-language model for objects present in image-caption pairs. The generated pseudo-mask annotations are then used to supervise an instance segmentation model, freeing the entire pipeline from any labour-expensive instance-level annotations and overfitting. Our extensive experiments show that our method trained with just pseudo-masks significantly improves the mAP scores on the MS-COCO dataset and OpenImages dataset compared to the recent state-of-the-art methods trained with manual masks. Codes and models are provided in https://vibashan.github.io/ovis-web/.
****
Complete-to-Partial 4D Distillation for Self-Supervised Point Cloud Sequence Representation Learning
Zhuoyang Zhang, Yuhao Dong, Yunze Liu, Li Yi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17661-17670
Recent work on 4D point cloud sequences has attracted a lot of attention. However, obtaining exhaustively labeled 4D datasets is often very expensive and laborious, so it is especially important to investigate how to utilize raw unlabeled data. However, most existing self-supervised point cloud representation learning methods only consider geometry from a static snapshot omitting the fact that sequential observations of dynamic scenes could reveal more comprehensive geometric details. To overcome such issues, this paper proposes a new 4D self-supervised pre-training method called Complete-to-Partial 4D Distillation. Our key idea is to formulate 4D self-supervised representation learning as a teacher-student knowledge distillation framework and let the student learn useful 4D representations with the guidance of the teacher. Experiments show that this approach significantly outperforms previous pre-training approaches on a wide range of 4D point cloud sequence understanding tasks. Code is available at: https://github.com/dongyh20/C2P.
****
BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects
Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, Stan Birchfield; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 606-617
We present a near real-time (10Hz) method for 6-DoF tracking of an unknown object from a monocular RGBD video sequence, while simultaneously performing neural 3

D reconstruction of the object. Our method works for arbitrary rigid objects, even when visual texture is largely absent. The object is assumed to be segmented in the first frame only. No additional information is required, and no assumption is made about the interaction agent. Key to our method is a Neural Object Field that is learned concurrently with a pose graph optimization process in order to robustly accumulate information into a consistent 3D representation capturing both geometry and appearance. A dynamic pool of posed memory frames is automatically maintained to facilitate communication between these threads. Our approach handles challenging sequences with large pose changes, partial and full occlusion, untextured surfaces, and specular highlights. We show results on HO3D, YCBInEOAT, and BEHAVE datasets, demonstrating that our method significantly outperforms existing approaches. Project page: https://bundlesdf.github.io/
****
Multi-Modal Gait Recognition via Effective Spatial-Temporal Feature Fusion
Yufeng Cui, Yimei Kang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17949-17957
Gait recognition is a biometric technology that identifies people by their walking patterns. The silhouettes-based method and the skeletons-based method are the two most popular approaches. However, the silhouette data are easily affected by clothing occlusion, and the skeleton data lack body shape information. To obtain a more robust and comprehensive gait representation for recognition, we propose a transformer-based gait recognition framework called MMGaitFormer, which effectively fuses and aggregates the spatial-temporal information from the skeletons and silhouettes. Specifically, a Spatial Fusion Module (SFM) and a Temporal Fusion Module (TFM) are proposed for effective spatial-level and temporal-level feature fusion, respectively. The SFM performs fine-grained body parts spatial fusion and guides the alignment of each part of the silhouette and each joint of the skeleton through the attention mechanism. The TFM performs temporal modeling through Cycle Position Embedding (CPE) and fuses temporal information of two modalities. Experiments demonstrate that our MMGaitFormer achieves state-of-the-art performance on popular gait datasets. For the most challenging "CL" (i.e., walking in different clothes) condition in CASIA-B, our method achieves a rank-1 accuracy of 94.8%, which outperforms the state-of-the-art single-modal methods by a large margin.
****
Crowd3D: Towards Hundreds of People Reconstruction From a Single Image
Hao Wen, Jing Huang, Huili Cui, Haozhe Lin, Yu-Kun Lai, Lu Fang, Kun Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8937-8946
Image-based multi-person reconstruction in wide-field large scenes is critical for crowd analysis and security alert. However, existing methods cannot deal with large scenes containing hundreds of people, which encounter the challenges of large number of people, large variations in human scale, and complex spatial distribution. In this paper, we propose Crowd3D, the first framework to reconstruct the 3D poses, shapes and locations of hundreds of people with global consistency from a single large-scene image. The core of our approach is to convert the problem of complex crowd localization into pixel localization with the help of our newly defined concept, Human-scene Virtual Interaction Point (HVIP). To reconstruct the crowd with global consistency, we propose a progressive reconstruction network based on HVIP by pre-estimating a scene-level camera and a ground plane. To deal with a large number of persons and various human sizes, we also design an adaptive human-centric cropping scheme. Besides, we contribute a benchmark dataset, LargeCrowd, for crowd reconstruction in a large scene. Experimental results demonstrate the effectiveness of the proposed method. The code and the dataset are available at http://cic.tju.edu.cn/faculty/likun/projects/Crowd3D.
****
Highly Confident Local Structure Based Consensus Graph Learning for Incomplete Multi-View Clustering
Jie Wen, Chengliang Liu, Gehui Xu, Zhihao Wu, Chao Huang, Lunke Fei, Yong Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition

(CVPR), 2023, pp. 15712-15721
Graph-based multi-view clustering has attracted extensive attention because of the powerful clustering-structure representation ability and noise robustness. Considering the reality of a large amount of incomplete data, in this paper, we propose a simple but effective method for incomplete multi-view clustering based on consensus graph learning, termed as HCLS_CGL. Unlike existing methods that utilize graph constructed from raw data to aid in the learning of consistent representation, our method directly learns a consensus graph across views for clustering. Specifically, we design a novel confidence graph and embed it to form a confidence structure driven consensus graph learning model. Our confidence graph is based on an intuitive similar-nearest-neighbor hypothesis, which does not require any additional information and can help the model to obtain a high-quality consensus graph for better clustering. Numerous experiments are performed to confirm the effectiveness of our method.
****
Humans As Light Bulbs: 3D Human Reconstruction From Thermal Reflection
Ruoshi Liu, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12531-12542
The relatively hot temperature of the human body causes people to turn into long-wave infrared light sources. Since this emitted light has a larger wavelength than visible light, many surfaces in typical scenes act as infrared mirrors with strong specular reflections. We exploit the thermal reflections of a person onto objects in order to locate their position and reconstruct their pose, even if they are not visible to a normal camera. We propose an analysis-by-synthesis framework that jointly models the objects, people, and their thermal reflections, which allows us to combine generative models with differentiable rendering of reflections. Quantitative and qualitative experiments show our approach works in highly challenging cases, such as with curved mirrors or when the person is completely unseen by a normal camera.
****
Hierarchical Discriminative Learning Improves Visual Representations of Biomedical Microscopy
Cheng Jiang, Xinhai Hou, Akhil Kondepudi, Asadur Chowdury, Christian W. Freudiger, Daniel A. Orringer, Honglak Lee, Todd C. Hollon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19798-19808
Learning high-quality, self-supervised, visual representations is essential to advance the role of computer vision in biomedical microscopy and clinical medicine. Previous work has focused on self-supervised representation learning (SSL) methods developed for instance discrimination and applied them directly to image patches, or fields-of-view, sampled from gigapixel whole-slide images (WSIs) used for cancer diagnosis. However, this strategy is limited because it (1) assumes patches from the same patient are independent, (2) neglects the patient-slide-patch hierarchy of clinical biomedical microscopy, and (3) requires strong data augmentations that can degrade downstream performance. Importantly, sampled patches from WSIs of a patient's tumor are a diverse set of image examples that capture the same underlying cancer diagnosis. This motivated HiDisc, a data-driven method that leverages the inherent patient-slide-patch hierarchy of clinical biomedical microscopy to define a hierarchical discriminative learning task that implicitly learns features of the underlying diagnosis. HiDisc uses a self-supervised contrastive learning framework in which positive patch pairs are defined based on a common ancestry in the data hierarchy, and a unified patch, slide, and patient discriminative learning objective is used for visual SSL. We benchmark HiDisc visual representations on two vision tasks using two biomedical microscopy datasets, and demonstrate that (1) HiDisc pretraining outperforms current state-of-the-art self-supervised pretraining methods for cancer diagnosis and genetic mutation prediction, and (2) HiDisc learns high-quality visual representations using natural patch diversity without strong data augmentations.
****
ProD: Prompting-To-Disentangle Domain Knowledge for Cross-Domain Few-Shot Image

Classification
Tianyi Ma, Yifan Sun, Zongxin Yang, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19754-19763
This paper considers few-shot image classification under the cross-domain scenario, where the train-to-test domain gap compromises classification accuracy. To mitigate the domain gap, we propose a prompting-to-disentangle (ProD) method through a novel exploration with the prompting mechanism. ProD adopts the popular multi-domain training scheme and extracts the backbone feature with a standard Convolutional Neural Network. Based on these two common practices, the key point of ProD is using the prompting mechanism in the transformer to disentangle the domain-general (DG) and domain-specific (DS) knowledge from the backbone feature. Specifically, ProD concatenates a DG and a DS prompt to the backbone feature and feeds them into a lightweight transformer. The DG prompt is learnable and shared by all the training domains, while the DS prompt is generated from the domain-of-interest on the fly. As a result, the transformer outputs DG and DS features in parallel with the two prompts, yielding the disentangling effect. We show that: 1) Simply sharing a single DG prompt for all the training domains already improves generalization towards the novel test domain. 2) The cross-domain generalization can be further reinforced by making the DG prompt neutral towards the training domains. 3) When inference, the DS prompt is generated from the support samples and can capture test domain knowledge through the prompting mechanism. Combining all three benefits, ProD significantly improves cross-domain few-shot classification. For instance, on CUB, ProD improves the 5-way 5-shot accuracy from 73.56% (baseline) to 79.19%, setting a new state of the art.
****
Clothing-Change Feature Augmentation for Person Re-Identification
Ke Han, Shaogang Gong, Yan Huang, Liang Wang, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22066-22075
Clothing-change person re-identification (CC Re-ID) aims to match the same person who changes clothes across cameras. Current methods are usually limited by the insufficient number and variation of clothing in training data, e.g. each person only has 2 outfits in the PRCC dataset. In this work, we propose a novel Clothing-Change Feature Augmentation (CCFA) model for CC Re-ID to largely expand clothing-change data in the feature space rather than visual image space. It automatically models the feature distribution expansion that reflects a person's clothing colour and texture variations to augment model training. Specifically, to formulate meaningful clothing variations in the feature space, our method first estimates a clothing-change normal distribution with intra-ID cross-clothing variances. Then an augmentation generator learns to follow the estimated distribution to augment plausible clothing-change features. The augmented features are guaranteed to maximise the change of clothing and minimise the change of identity properties by adversarial learning to assure the effectiveness. Such augmentation is performed iteratively with an ID-correlated augmentation strategy to increase intra-ID clothing variations and reduce inter-ID clothing variations, enforcing the Re-ID model to learn clothing-independent features inherently. Extensive experiments demonstrate the effectiveness of our method with state-of-the-art results on CC Re-ID datasets.
****
CafeBoost: Causal Feature Boost To Eliminate Task-Induced Bias for Class Incremental Learning
Benliu Qiu, Hongliang Li, Haitao Wen, Heqian Qiu, Lanxiao Wang, Fanman Meng, Qingbo Wu, Lili Pan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16016-16025
Continual learning requires a model to incrementally learn a sequence of tasks and aims to predict well on all the learned tasks so far, which notoriously suffers from the catastrophic forgetting problem. In this paper, we find a new type of bias appearing in continual learning, coined as task-induced bias. We place continual learning into a causal framework, based on which we find the task-induced bias is reduced naturally by two underlying mechanisms in task and domain incr

emental learning. However, these mechanisms do not exist in class incremental learning (CIL), in which each task contains a unique subset of classes. To eliminate the task-induced bias in CIL, we devise a causal intervention operation so as to cut off the causal path that causes the task-induced bias, and then implement it as a causal debias module that transforms biased features into unbiased ones. In addition, we propose a training pipeline to incorporate the novel module into existing methods and jointly optimize the entire architecture. Our overall approach does not rely on data replay, and is simple and convenient to plug into existing methods. Extensive empirical study on CIFAR-100 and ImageNet shows that our approach can improve accuracy and reduce forgetting of well-established methods by a large margin.
****
A-La-Carte Prompt Tuning (APT): Combining Distinct Data via Composable Prompting
Benjamin Bowman, Alessandro Achille, Luca Zancato, Matthew Trager, Pramuditha Perera, Giovanni Paolini, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14984-14993
We introduce A-la-carte Prompt Tuning (APT), a transformer-based scheme to tune prompts on distinct data so that they can be arbitrarily composed at inference time. The individual prompts can be trained in isolation, possibly on different devices, at different times, and on different distributions or domains. Furthermore each prompt only contains information about the subset of data it was exposed to during training. During inference, models can be assembled based on arbitrary selections of data sources, which we call a-la-carte learning. A-la-carte learning enables constructing bespoke models specific to each user's individual access rights and preferences. We can add or remove information from the model by simply adding or removing the corresponding prompts without retraining from scratch. We demonstrate that a-la-carte built models achieve accuracy within 5% of models trained on the union of the respective sources, with comparable cost in terms of training and inference time. For the continual learning benchmarks Split CIFAR-100 and CORe50, we achieve state-of-the-art performance.
****
ImageNet-E: Benchmarking Neural Network Robustness via Attribute Editing
Xiaodan Li, Yuefeng Chen, Yao Zhu, Shuhui Wang, Rong Zhang, Hui Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20371-20381
Recent studies have shown that higher accuracy on ImageNet usually leads to better robustness against different corruptions. In this paper, instead of following the traditional research paradigm that investigates new out-of-distribution corruptions or perturbations deep models may encounter, we conduct model debugging in in-distribution data to explore which object attributes a model may be sensitive to. To achieve this goal, we create a toolkit for object editing with controls of backgrounds, sizes, positions, and directions, and create a rigorous benchmark named ImageNet-E(diting) for evaluating the image classifier robustness in terms of object attributes. With our ImageNet-E, we evaluate the performance of current deep learning models, including both convolutional neural networks and vision transformers. We find that most models are quite sensitive to attribute changes. An imperceptible change in the background can lead to an average of 9.23% drop on top-1 accuracy. We also evaluate some robust models including both adversarially trained models and other robust trained models and find that some models show worse robustness against attribute changes than vanilla models. Based on these findings, we discover ways to enhance attribute robustness with preprocessing, architecture designs, and training strategies. We hope this work can provide some insights to the community and open up a new avenue for research in robust computer vision. The code and dataset will be publicly available.
****
Learning With Fantasy: Semantic-Aware Virtual Contrastive Constraint for Few-Shot Class-Incremental Learning
Zeyin Song, Yifan Zhao, Yujun Shi, Peixi Peng, Li Yuan, Yonghong Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24183-24192

Few-shot class-incremental learning (FSCIL) aims at learning to classify new classes continually from limited samples without forgetting the old classes. The mainstream framework tackling FSCIL is first to adopt the cross-entropy (CE) loss for training at the base session, then freeze the feature extractor to adapt to new classes. However, in this work, we find that the CE loss is not ideal for the base session training as it suffers poor class separation in terms of representations, which further degrades generalization to novel classes. One tempting method to mitigate this problem is to apply an additional naive supervised contrastive learning (SCL) in the base session. Unfortunately, we find that although SCL can create a slightly better representation separation among different base classes, it still struggles to separate base classes and new classes. Inspired by the observations made, we propose Semantic-Aware Virtual Contrastive model (SAVC), a novel method that facilitates separation between new classes and base classes by introducing virtual classes to SCL. These virtual classes, which are generated via pre-defined transformations, not only act as placeholders for unseen classes in the representation space but also provide diverse semantic information. By learning to recognize and contrast in the fantasy space fostered by virtual classes, our SAVC significantly boosts base class separation and novel class generalization, achieving new state-of-the-art performance on the three widely-used FSCIL benchmark datasets. Code is available at: https://github.com/zysong0113/SAVC.
****

ViLEM: Visual-Language Error Modeling for Image-Text Retrieval

Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Ying Shan, Bing Li, Weiming Hu, Xiaohu Qie, Jianping Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11018-11027

Dominant pre-training works for image-text retrieval adopt "dual-encoder" architecture to enable high efficiency, where two encoders are used to extract image and text representations and contrastive learning is employed for global alignment. However, coarse-grained global alignment ignores detailed semantic associations between image and text. In this work, we propose a novel proxy task, named Visual-Language Error Modeling (ViLEM), to inject detailed image-text association into "dual-encoder" model by "proofreading" each word in the text against the corresponding image. Specifically, we first edit the image-paired text to automatically generate diverse plausible negative texts with pre-trained language models. ViLEM then enforces the model to discriminate the correctness of each word in the plausible negative texts and further correct the wrong words via resorting to image information. Furthermore, we propose a multi-granularity interaction framework to perform ViLEM via interacting text features with both global and local image features, which associates local text semantics with both high-level visual context and multi-level local visual information. Our method surpasses state-of-the-art "dual-encoder" methods by a large margin on the image-text retrieval task and significantly improves discriminativeness to local textual semantics. Our model can also generalize well to video-text retrieval.
****

Egocentric Auditory Attention Localization in Conversations

Fiona Ryan, Hao Jiang, Abhinav Shukla, James M. Rehg, Vamsi Krishna Ithapu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14663-14674

In a noisy conversation environment such as a dinner party, people often exhibit selective auditory attention, or the ability to focus on a particular speaker while tuning out others. Recognizing who somebody is listening to in a conversation is essential for developing technologies that can understand social behavior and devices that can augment human hearing by amplifying particular sound sources. The computer vision and audio research communities have made great strides towards recognizing sound sources and speakers in scenes. In this work, we take a step further by focusing on the problem of localizing auditory attention targets in egocentric video, or detecting who in a camera wearer's field of view they are listening to. To tackle the new and challenging Selective Auditory Attention Localization problem, we propose an end-to-end deep learning approach that uses

egocentric video and multichannel audio to predict the heatmap of the camera wea rer's auditory attention. Our approach leverages spatiotemporal audiovisual feat ures and holistic reasoning about the scene to make predictions, and outperforms a set of baselines on a challenging multi-speaker conversation dataset. Project page: https://fkryan.github.io/saal
****
Texture-Guided Saliency Distilling for Unsupervised Salient Object Detection
Huajun Zhou, Bo Qiao, Lingxiao Yang, Jianhuang Lai, Xiaohua Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7257-7267
Deep Learning-based Unsupervised Salient Object Detection (USOD) mainly relies o n the noisy saliency pseudo labels that have been generated from traditional han dcraft methods or pre-trained networks. To cope with the noisy labels problem, a class of methods focus on only easy samples with reliable labels but ignore val uable knowledge in hard samples. In this paper, we propose a novel USOD method t o mine rich and accurate saliency knowledge from both easy and hard samples. Fir st, we propose a Confidence-aware Saliency Distilling (CSD) strategy that scores samples conditioned on samples' confidences, which guides the model to distill saliency knowledge from easy samples to hard samples progressively. Second, we p ropose a Boundary-aware Texture Matching (BTM) strategy to refine the boundaries of noisy labels by matching the textures around the predicted boundaries. Exten sive experiments on RGB, RGB-D, RGB-T, and video SOD benchmarks prove that our m ethod achieves state-of-the-art USOD performance. Code is available at www.githu b.com/moothes/A2S-v2.
****
AltFreezing for More General Video Face Forgery Detection
Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Houqiang Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 23, pp. 4129-4138
Existing face forgery detection models try to discriminate fake images by detect ing only spatial artifacts (e.g., generative artifacts, blending) or mainly temp oral artifacts (e.g., flickering, discontinuity). They may experience significan t performance degradation when facing out-domain artifacts. In this paper, we pr opose to capture both spatial and temporal artifacts in one model for face forge ry detection. A simple idea is to leverage a spatiotemporal model (3D ConvNet). However, we find that it may easily rely on one type of artifact and ignore the other. To address this issue, we present a novel training strategy called AltFre ezing for more general face forgery detection. The AltFreezing aims to encourage the model to detect both spatial and temporal artifacts. It divides the weights of a spatiotemporal network into two groups: spatial- and temporal-related. The n the two groups of weights are alternately frozen during the training process s o that the model can learn spatial and temporal features to distinguish real or fake videos. Furthermore, we introduce various video-level data augmentation met hods to improve the generalization capability of the forgery detection model. Ex tensive experiments show that our framework outperforms existing methods in term s of generalization to unseen manipulations and datasets.
****
Cascaded Local Implicit Transformer for Arbitrary-Scale Super-Resolution
Hao-Wei Chen, Yu-Syuan Xu, Min-Fong Hong, Yi-Min Tsai, Hsien-Kai Kuo, Chun-Yi Le e; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni tion (CVPR), 2023, pp. 18257-18267
Implicit neural representation demonstrates promising ability in representing im ages with arbitrary resolutions recently. In this paper, we present Local Implic it Transformer (LIT) that integrates attention mechanism and frequency encoding technique into local implicit image function. We design a cross-scale local atte ntion block to effectively aggregate local features and a local frequency encodi ng block to combine positional encoding with Fourier domain information for cons tructing high-resolution (HR) images. To further improve representative power, w e propose Cascaded LIT (CLIT) exploiting multi-scale features along with cumulat ive training strategy that gradually increase the upsampling factors for trainin

g. We have performed extensive experiments to validate the effectiveness of thes
e components and analyze the variants of the training strategy. The qualitative
and quantitative results demonstrated that LIT and CLIT achieve favorable result
s and outperform the previous works within arbitrary super-resolution tasks.
****
Learning Partial Correlation Based Deep Visual Representation for Image Classifi
cation
Saimunur Rahman, Piotr Koniusz, Lei Wang, Luping Zhou, Peyman Moghadam, Changmin
g Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2023, pp. 6231-6240
Visual representation based on covariance matrix has demonstrates its efficacy f
or image classification by characterising the pairwise correlation of different
channels in convolutional feature maps. However, pairwise correlation will becom
e misleading once there is another channel correlating with both channels of int
erest, resulting in the "confounding" effect. For this case, "partial correlatio
n" which removes the confounding effect shall be estimated instead. Nevertheless
, reliably estimating partial correlation requires to solve a symmetric positive
 definite matrix optimisation, known as sparse inverse covariance estimation (SI
CE). How to incorporate this process into CNN remains an open issue. In this wor
k, we formulate SICE as a novel structured layer of CNN. To ensure end-to-end tr
ainability, we develop an iterative method to solve the above matrix optimisatio
n during forward and backward propagation steps. Our work obtains a partial corr
elation based deep visual representation and mitigates the small sample problem
often encountered by covariance matrix estimation in CNN. Computationally, our m
odel can be effectively trained with GPU and works well with a large number of c
hannels of advanced CNNs. Experiments show the efficacy and superior classificat
ion performance of our deep visual representation compared to covariance matrix
based counterparts.
****
Open-World Multi-Task Control Through Goal-Aware Representation Learning and Ada
ptive Horizon Prediction
Shaofei Cai, Zihao Wang, Xiaojian Ma, Anji Liu, Yitao Liang; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
 13734-13744
We study the problem of learning goal-conditioned policies in Minecraft, a popul
ar, widely accessible yet challenging open-ended environment for developing huma
n-level multi-task agents. We first identify two main challenges of learning suc
h policies: 1) the indistinguishability of tasks from the state distribution, du
e to the vast scene diversity, and 2) the non-stationary nature of environment d
ynamics caused by the partial observability. To tackle the first challenge, we p
ropose Goal-Sensitive Backbone (GSB) for the policy to encourage the emergence o
f goal-relevant visual state representations. To tackle the second challenge, th
e policy is further fueled by an adaptive horizon prediction module that helps a
lleviate the learning uncertainty brought by the non-stationary dynamics. Experi
ments on 20 Minecraft tasks show that our method significantly outperforms the b
est baseline so far; in many of them, we double the performance. Our ablation an
d exploratory studies then explain how our approach beat the counterparts and al
so unveil the surprising bonus of zero-shot generalization to new scenes (biomes
). We hope our agent could help shed some light on learning goal-conditioned, mu
lti-task agents in challenging, open-ended environments like Minecraft.
****
MoDi: Unconditional Motion Synthesis From Diverse Data
Sigal Raab, Inbal Leibovitch, Peizhuo Li, Kfir Aberman, Olga Sorkine-Hornung, Da
niel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat
tern Recognition (CVPR), 2023, pp. 13873-13883
The emergence of neural networks has revolutionized the field of motion synthesi
s. Yet, learning to unconditionally synthesize motions from a given distribution
 remains challenging, especially when the motions are highly diverse. In this wo
rk, we present MoDi -- a generative model trained in an unsupervised setting fro
m an extremely diverse, unstructured and unlabeled dataset. During inference, Mo

Di can synthesize high-quality, diverse motions. Despite the lack of any structure in the dataset, our model yields a well-behaved and highly structured latent space, which can be semantically clustered, constituting a strong motion prior that facilitates various applications including semantic editing and crowd animation. In addition, we present an encoder that inverts real motions into MoDi's natural motion manifold, issuing solutions to various ill-posed challenges such as completion from prefix and spatial editing. Our qualitative and quantitative experiments achieve state-of-the-art results that outperform recent SOTA techniques. Code and trained models are available at https://sigal-raab.github.io/MoDi.
****
Visual Localization Using Imperfect 3D Models From the Internet
Vojtech Panek, Zuzana Kukelova, Torsten Sattler; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13175-13186
Visual localization is a core component in many applications, including augmented reality (AR). Localization algorithms compute the camera pose of a query image w.r.t. a scene representation, which is typically built from images. This often requires capturing and storing large amounts of data, followed by running Structure-from-Motion (SfM) algorithms. An interesting, and underexplored, source of data for building scene representations are 3D models that are readily available on the Internet, e.g., hand-drawn CAD models, 3D models generated from building footprints, or from aerial images. These models allow to perform visual localization right away without the time-consuming scene capturing and model building steps. Yet, it also comes with challenges as the available 3D models are often imperfect reflections of reality. E.g., the models might only have generic or no textures at all, might only provide a simple approximation of the scene geometry, or might be stretched. This paper studies how the imperfections of these models affect localization accuracy. We create a new benchmark for this task and provide a detailed experimental evaluation based on multiple 3D models per scene. We show that 3D models from the Internet show promise as an easy-to-obtain scene representation. At the same time, there is significant room for improvement for visual localization pipelines. To foster research on this interesting and challenging task, we release our benchmark at v-pnk.github.io/cadloc.
****
Network-Free, Unsupervised Semantic Segmentation With Synthetic Images
Qianli Feng, Raghudeep Gadde, Wentong Liao, Eduard Ramon, Aleix Martinez; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23602-23610
We derive a method that yields highly accurate semantic segmentation maps without the use of any additional neural network, layers, manually annotated training data, or supervised training. Our method is based on the observation that the correlation of a set of pixels belonging to the same semantic segment do not change when generating synthetic variants of an image using the style mixing approach in GANs. We show how we can use GAN inversion to accurately semantically segment synthetic and real photos as well as generate large training image-semantic segmentation mask pairs for downstream tasks.
****
Hierarchical Dense Correlation Distillation for Few-Shot Segmentation
Bohao Peng, Zhuotao Tian, Xiaoyang Wu, Chengyao Wang, Shu Liu, Jingyong Su, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23641-23651
Few-shot semantic segmentation (FSS) aims to form class-agnostic models segmenting unseen classes with only a handful of annotations. Previous methods limited to the semantic feature and prototype representation suffer from coarse segmentation granularity and train-set overfitting. In this work, we design Hierarchically Decoupled Matching Network (HDMNet) mining pixel-level support correlation based on the transformer architecture. The self-attention modules are used to assist in establishing hierarchical dense features, as a means to accomplish the cascade matching between query and support features. Moreover, we propose a matching module to reduce train-set overfitting and introduce correlation distillation leveraging semantic correspondence from coarse resolution to boost fine-grained s

egmentation. Our method performs decently in experiments. We achieve 50.0% mIoU on COCO-5i dataset one-shot setting and 56.0% on five-shot segmentation, respectively. The code is available on the project website.
****
PVO: Panoptic Visual Odometry
Weicai Ye, Xinyue Lan, Shuo Chen, Yuhang Ming, Xingyuan Yu, Hujun Bao, Zhaopeng Cui, Guofeng Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9579-9589
We present PVO, a novel panoptic visual odometry framework to achieve more comprehensive modeling of the scene motion, geometry, and panoptic segmentation information. Our PVO models visual odometry (VO) and video panoptic segmentation (VPS) in a unified view, which makes the two tasks mutually beneficial. Specifically, we introduce a panoptic update module into the VO Module with the guidance of image panoptic segmentation. This Panoptic-Enhanced VO Module can alleviate the impact of dynamic objects in the camera pose estimation with a panoptic-aware dynamic mask. On the other hand, the VO-Enhanced VPS Module also improves the segmentation accuracy by fusing the panoptic segmentation result of the current frame on the fly to the adjacent frames, using geometric information such as camera pose, depth, and optical flow obtained from the VO Module. These two modules contribute to each other through recurrent iterative optimization. Extensive experiments demonstrate that PVO outperforms state-of-the-art methods in both visual odometry and video panoptic segmentation tasks.
****
Generative Diffusion Prior for Unified Image Restoration and Enhancement
Ben Fei, Zhaoyang Lyu, Liang Pan, Junzhe Zhang, Weidong Yang, Tianyue Luo, Bo Zhang, Bo Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9935-9946
Existing image restoration methods mostly leverage the posterior distribution of natural images. However, they often assume known degradation and also require supervised training, which restricts their adaptation to complex real applications. In this work, we propose the Generative Diffusion Prior (GDP) to effectively model the posterior distributions in an unsupervised sampling manner. GDP utilizes a pre-train denoising diffusion generative model (DDPM) for solving linear inverse, non-linear, or blind problems. Specifically, GDP systematically explores a protocol of conditional guidance, which is verified more practical than the commonly used guidance way. Furthermore, GDP is strength at optimizing the parameters of degradation model during denoising process, achieving blind image restoration. Besides, we devise hierarchical guidance and patch-based methods, enabling the GDP to generate images of arbitrary resolutions. Experimentally, we demonstrate GDP's versatility on several image datasets for linear problems, such as super-resolution, deblurring, inpainting, and colorization, as well as non-linear and blind issues, such as low-light enhancement and HDR image recovery. GDP outperforms the current leading unsupervised methods on the diverse benchmarks in reconstruction quality and perceptual quality. Moreover, GDP also generalizes well for natural images or synthesized images with arbitrary sizes from various tasks out of the distribution of the ImageNet training set.
****
Real-Time Controllable Denoising for Image and Video
Zhaoyang Zhang, Yitong Jiang, Wenqi Shao, Xiaogang Wang, Ping Luo, Kaimo Lin, Jinwei Gu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14028-14038
Controllable image denoising aims to generate clean samples with human perceptual priors and balance sharpness and smoothness. In traditional filter-based denoising methods, this can be easily achieved by adjusting the filtering strength. However, for NN (Neural Network)-based models, adjusting the final denoising strength requires performing network inference each time, making it almost impossible for real-time user interaction. In this paper, we introduce Real-time Controllable Denoising (RCD), the first deep image and video denoising pipeline that provides a fully controllable user interface to edit arbitrary denoising levels in real-time with only one-time network inference. Unlike existing controllable den

oising methods that require multiple denoisers and training stages, RCD replaces the last output layer (which usually outputs a single noise map) of an existing CNN-based model with a lightweight module that outputs multiple noise maps. We propose a novel Noise Decorrelation process to enforce the orthogonality of the noise feature maps, allowing arbitrary noise level control through noise map interpolation. This process is network-free and does not require network inference. Our experiments show that RCD can enable real-time editable image and video denoising for various existing heavy-weight models without sacrificing their original performance.
****
ISBNet: A 3D Point Cloud Instance Segmentation Network With Instance-Aware Sampling and Box-Aware Dynamic Convolution
Tuan Duc Ngo, Binh-Son Hua, Khoi Nguyen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13550-13559
Existing 3D instance segmentation methods are predominated by the bottom-up design -- manually fine-tuned algorithm to group points into clusters followed by a refinement network. However, by relying on the quality of the clusters, these methods generate susceptible results when (1) nearby objects with the same semantic class are packed together, or (2) large objects with loosely connected regions. To address these limitations, we introduce ISBNet, a novel cluster-free method that represents instances as kernels and decodes instance masks via dynamic convolution. To efficiently generate high-recall and discriminative kernels, we propose a simple strategy named Instance-aware Farthest Point Sampling to sample candidates and leverage the local aggregation layer inspired by PointNet++ to encode candidate features. Moreover, we show that predicting and leveraging the 3D axis-aligned bounding boxes in the dynamic convolution further boosts performance. Our method set new state-of-the-art results on ScanNetV2 (55.9), S3DIS (60.8), and STPLS3D (49.2) in terms of AP and retains fast inference time (237ms per scene on ScanNetV2). The source code and trained models are available at https://github.com/VinAIResearch/ISBNet.
****
Hi4D: 4D Instance Segmentation of Close Human Interaction
Yifei Yin, Chen Guo, Manuel Kaufmann, Juan Jose Zarate, Jie Song, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17016-17027
We propose Hi4D, a method and dataset for the auto analysis of physically close human-human interaction under prolonged contact. Robustly disentangling several in-contact subjects is a challenging task due to occlusions and complex shapes. Hence, existing multi-view systems typically fuse 3D surfaces of close subjects into a single, connected mesh. To address this issue we leverage i) individually fitted neural implicit avatars; ii) an alternating optimization scheme that refines pose and surface through periods of close proximity; and iii) thus segment the fused raw scans into individual instances. From these instances we compile Hi4D dataset of 4D textured scans of 20 subject pairs, 100 sequences, and a total of more than 11K frames. Hi4D contains rich interaction-centric annotations in 2D and 3D alongside accurately registered parametric body models. We define varied human pose and shape estimation tasks on this dataset and provide results from state-of-the-art methods on these benchmarks.
****
Hi-LASSIE: High-Fidelity Articulated Shape and Skeleton Discovery From Sparse Image Ensemble
Chun-Han Yao, Wei-Chih Hung, Yuanzhen Li, Michael Rubinstein, Ming-Hsuan Yang, Varun Jampani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4853-4862
Automatically estimating 3D skeleton, shape, camera viewpoints, and part articulation from sparse in-the-wild image ensembles is a severely under-constrained and challenging problem. Most prior methods rely on large-scale image datasets, dense temporal correspondence, or human annotations like camera pose, 2D keypoints, and shape templates. We propose Hi-LASSIE, which performs 3D articulated reconstruction from only 20-30 online images in the wild without any user-defined sha

pe or skeleton templates. We follow the recent work of LASSIE that tackles a sim ilar problem setting and make two significant advances. First, instead of relyin g on a manually annotated 3D skeleton, we automatically estimate a class-specifi c skeleton from the selected reference image. Second, we improve the shape recon structions with novel instance-specific optimization strategies that allow recon structions to faithful fit on each instance while preserving the class-specific priors learned across all images. Experiments on in-the-wild image ensembles sho w that Hi-LASSIE obtains higher fidelity state-of-the-art 3D reconstructions des pite requiring minimum user input. Project page: chhankyao.github.io/hi-lassie/
****
IterativePFN: True Iterative Point Cloud Filtering
Dasith de Silva Edirimuni, Xuequan Lu, Zhiwen Shao, Gang Li, Antonio Robles-Kell y, Ying He; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter n Recognition (CVPR), 2023, pp. 13530-13539
The quality of point clouds is often limited by noise introduced during their ca pture process. Consequently, a fundamental 3D vision task is the removal of nois e, known as point cloud filtering or denoising. State-of-the-art learning based methods focus on training neural networks to infer filtered displacements and di rectly shift noisy points onto the underlying clean surfaces. In high noise cond itions, they iterate the filtering process. However, this iterative filtering is only done at test time and is less effective at ensuring points converge quickl y onto the clean surfaces. We propose IterativePFN (iterative point cloud filter ing network), which consists of multiple IterationModules that model the true it erative filtering process internally, within a single network. We train our Iter ativePFN network using a novel loss function that utilizes an adaptive ground tr uth target at each iteration to capture the relationship between intermediate fi ltering results during training. This ensures that the filtered results converge faster to the clean surfaces. Our method is able to obtain better performance c ompared to state-of-the-art methods. The source code can be found at: https://gi thub.com/ddsediri/IterativePFN.
****
Computationally Budgeted Continual Learning: What Does Matter?
Ameya Prabhu, Hasan Abed Al Kader Hammoud, Puneet K. Dokania, Philip H.S. Torr, Ser-Nam Lim, Bernard Ghanem, Adel Bibi; Proceedings of the IEEE/CVF Conference o n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3698-3707
Continual Learning (CL) aims to sequentially train models on streams of incoming data that vary in distribution by preserving previous knowledge while adapting to new data. Current CL literature focuses on restricted access to previously se en data, while imposing no constraints on the computational budget for training. This is unreasonable for applications in-the-wild, where systems are primarily constrained by computational and time budgets, not storage. We revisit this prob lem with a large-scale benchmark and analyze the performance of traditional CL a pproaches in a compute-constrained setting, where effective memory samples used in training can be implicitly restricted as a consequence of limited computation . We conduct experiments evaluating various CL sampling strategies, distillation losses, and partial fine-tuning on two large-scale datasets, namely ImageNet2K and Continual Google Landmarks V2 in data incremental, class incremental, and ti me incremental settings. Through extensive experiments amounting to a total of o ver 1500 GPU-hours, we find that, under compute-constrained setting, traditional CL approaches, with no exception, fail to outperform a simple minimal baseline that samples uniformly from memory. Our conclusions are consistent in a differen t number of stream time steps, e.g., 20 to 200, and under several computational budgets. This suggests that most existing CL methods are particularly too comput ationally expensive for realistic budgeted deployment. Code for this project is available at: https://github.com/drimpossible/BudgetCL.
****
Decentralized Learning With Multi-Headed Distillation
Andrey Zhmoginov, Mark Sandler, Nolan Miller, Gus Kristiansen, Max Vladymyrov; P roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8053-8063

Decentralized learning with private data is a central problem in machine learnin g. We propose a novel distillation-based decentralized learning technique that a llows multiple agents with private non-iid data to learn from each other, withou t having to share their data, weights or weight updates. Our approach is communi cation efficient, utilizes an unlabeled public dataset and uses multiple auxilia ry heads for each client, greatly improving training efficiency in the case of h eterogeneous data. This approach allows individual models to preserve and enhanc e performance on their private tasks while also dramatically improving their per formance on the global aggregated data distribution. We study the effects of dat a and model architecture heterogeneity and the impact of the underlying communic ation graph topology on learning efficiency and show that our agents can signifi cantly improve their performance compared to learning in isolation.
****
SQUID: Deep Feature In-Painting for Unsupervised Anomaly Detection
Tiange Xiang, Yixiao Zhang, Yongyi Lu, Alan L. Yuille, Chaoyi Zhang, Weidong Cai , Zongwei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2023, pp. 23890-23901
Radiography imaging protocols focus on particular body regions, therefore produc ing images of great similarity and yielding recurrent anatomical structures acro ss patients. To exploit this structured information, we propose the use of Space -aware Memory Queues for In-painting and Detecting anomalies from radiography im ages (abbreviated as SQUID). We show that SQUID can taxonomize the ingrained ana tomical structures into recurrent patterns; and in the inference, it can identif y anomalies (unseen/modified patterns) in the image. SQUID surpasses 13 state-of -the-art methods in unsupervised anomaly detection by at least 5 points on two c hest X-ray benchmark datasets measured by the Area Under the Curve (AUC). Additi onally, we have created a new dataset (DigitAnatomy), which synthesizes the spat ial correlation and consistent shape in chest anatomy. We hope DigitAnatomy can prompt the development, evaluation, and interpretability of anomaly detection me thods.
****
CF-Font: Content Fusion for Few-Shot Font Generation
Chi Wang, Min Zhou, Tiezheng Ge, Yuning Jiang, Hujun Bao, Weiwei Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 023, pp. 1858-1867
Content and style disentanglement is an effective way to achieve few-shot font g eneration. It allows to transfer the style of the font image in a source domain to the style defined with a few reference images in a target domain. However, th e content feature extracted using a representative font might not be optimal. In light of this, we propose a content fusion module (CFM) to project the content feature into a linear space defined by the content features of basis fonts, whic h can take the variation of content features caused by different fonts into cons ideration. Our method also allows to optimize the style representation vector of reference images through a lightweight iterative style-vector refinement (ISR) strategy. Moreover, we treat the 1D projection of a character image as a probabi lity distribution and leverage the distance between two distributions as the rec onstruction loss (namely projected character loss, PCL). Compared to L2 or L1 re construction loss, the distribution distance pays more attention to the global s hape of characters. We have evaluated our method on a dataset of 300 fonts with 6.5k characters each. Experimental results verify that our method outperforms ex isting state-of-the-art few-shot font generation methods by a large margin. The source code can be found at https://github.com/wangchi95/CF-Font.
****
On the Convergence of IRLS and Its Variants in Outlier-Robust Estimation
Liangzu Peng, Christian Kümmerle, René Vidal; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17808-17818
Outlier-robust estimation involves estimating some parameters (e.g., 3D rotation s) from data samples in the presence of outliers, and is typically formulated as a non-convex and non-smooth problem. For this problem, the classical method cal led iteratively reweighted least-squares (IRLS) and its variants have shown impr

essive performance. This paper makes several contributions towards understanding why these algorithms work so well. First, we incorporate majorization and graduated non-convexity (GNC) into the IRLS framework and prove that the resulting IRLS variant is a convergent method for outlier-robust estimation. Moreover, in the robust regression context with a constant fraction of outliers, we prove this IRLS variant converges to the ground truth at a global linear and local quadratic rate for a random Gaussian feature matrix with high probability. Experiments corroborate our theory and show that the proposed IRLS variant converges within 5-10 iterations for typical problem instances of outlier-robust estimation, while state-of-the-art methods need at least 30 iterations. A basic implementation of our method is provided: https://github.com/liangzu/IRLS-CVPR2023
****
CLIP-S4: Language-Guided Self-Supervised Semantic Segmentation
Wenbin He, Suphanut Jamonnak, Liang Gou, Liu Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11207-11216
Existing semantic segmentation approaches are often limited by costly pixel-wise annotations and predefined classes. In this work, we present CLIP-S^4 that leverages self-supervised pixel representation learning and vision-language models to enable various semantic segmentation tasks (e.g., unsupervised, transfer learning, language-driven segmentation) without any human annotations and unknown class information. We first learn pixel embeddings with pixel-segment contrastive learning from different augmented views of images. To further improve the pixel embeddings and enable language-driven semantic segmentation, we design two types of consistency guided by vision-language models: 1) embedding consistency, aligning our pixel embeddings to the joint feature space of a pre-trained vision-language model, CLIP; and 2) semantic consistency, forcing our model to make the same predictions as CLIP over a set of carefully designed target classes with both known and unknown prototypes. Thus, CLIP-S^4 enables a new task of class-free semantic segmentation where no unknown class information is needed during training. As a result, our approach shows consistent and substantial performance improvement over four popular benchmarks compared with the state-of-the-art unsupervised and language-driven semantic segmentation methods. More importantly, our method outperforms these methods on unknown class recognition by a large margin.
****
Deep Incomplete Multi-View Clustering With Cross-View Partial Sample and Prototype Alignment
Jiaqi Jin, Siwei Wang, Zhibin Dong, Xinwang Liu, En Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11600-11609
The success of existing multi-view clustering relies on the assumption of sample integrity across multiple views. However, in real-world scenarios, samples of multi-view are partially available due to data corruption or sensor failure, which leads to incomplete multi-view clustering study (IMVC). Although several attempts have been proposed to address IMVC, they suffer from the following drawbacks: i) Existing methods mainly adopt cross-view contrastive learning forcing the representations of each sample across views to be exactly the same, which might ignore view discrepancy and flexibility in representations; ii) Due to the absence of non-observed samples across multiple views, the obtained prototypes of clusters might be unaligned and biased, leading to incorrect fusion. To address the above issues, we propose a Cross-view Partial Sample and Prototype Alignment Network (CPSPAN) for Deep Incomplete Multi-view Clustering. Firstly, unlike existing contrastive-based methods, we adopt pair-observed data alignment as 'proxy supervised signals' to guide instance-to-instance correspondence construction among views. Then, regarding of the shifted prototypes in IMVC, we further propose a prototype alignment module to achieve incomplete distribution calibration across views. Extensive experimental results showcase the effectiveness of our proposed modules, attaining noteworthy performance improvements when compared to existing IMVC competitors on benchmark datasets.
****

A New Comprehensive Benchmark for Semi-Supervised Video Anomaly Detection and Anticipation
Congqi Cao, Yue Lu, Peng Wang, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20392-20401
Semi-supervised video anomaly detection (VAD) is a critical task in the intelligent surveillance system. However, an essential type of anomaly in VAD named scene-dependent anomaly has not received the attention of researchers. Moreover, there is no research investigating anomaly anticipation, a more significant task for preventing the occurrence of anomalous events. To this end, we propose a new comprehensive dataset, NWPU Campus, containing 43 scenes, 28 classes of abnormal events, and 16 hours of videos. At present, it is the largest semi-supervised VAD dataset with the largest number of scenes and classes of anomalies, the longest duration, and the only one considering the scene-dependent anomaly. Meanwhile, it is also the first dataset proposed for video anomaly anticipation. We further propose a novel model capable of detecting and anticipating anomalous events simultaneously. Compared with 7 outstanding VAD algorithms in recent years, our method can cope with scene-dependent anomaly detection and anomaly anticipation both well, achieving state-of-the-art performance on ShanghaiTech, CUHK Avenue, IITB Corridor and the newly proposed NWPU Campus datasets consistently. Our dataset and code is available at: https://campusvad.github.io.
****
Revisiting Multimodal Representation in Contrastive Learning: From Patch and Token Embeddings to Finite Discrete Tokens
Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N. Metaxas, Hongxia Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15095-15104
Contrastive learning-based vision-language pre-training approaches, such as CLIP, have demonstrated great success in many vision-language tasks. These methods achieve cross-modal alignment by encoding a matched image-text pair with similar feature embeddings, which are generated by aggregating information from visual patches and language tokens. However, direct aligning cross-modal information using such representations is challenging, as visual patches and text tokens differ in semantic levels and granularities. To alleviate this issue, we propose a Finite Discrete Tokens (FDT) based multimodal representation. FDT is a set of learnable tokens representing certain visual-semantic concepts. Both images and texts are embedded using shared FDT by first grounding multimodal inputs to FDT space and then aggregating the activated FDT representations. The matched visual and semantic concepts are enforced to be represented by the same set of discrete tokens by a sparse activation constraint. As a result, the granularity gap between the two modalities is reduced. Through both quantitative and qualitative analyses, we demonstrate that using FDT representations in CLIP-style models improves cross-modal alignment and performance in visual recognition and vision-language downstream tasks. Furthermore, we show that our method can learn more comprehensive representations, and the learned FDT capture meaningful cross-modal correspondence, ranging from objects to actions and attributes.
****
3Mformer: Multi-Order Multi-Mode Transformer for Skeletal Action Recognition
Lei Wang, Piotr Koniusz; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5620-5631
Many skeletal action recognition models use GCNs to represent the human body by 3D body joints connected body parts. GCNs aggregate one- or few-hop graph neighbourhoods, and ignore the dependency between not linked body joints. We propose to form hypergraph to model hyper-edges between graph nodes (e.g., third- and fourth-order hyper-edges capture three and four nodes) which help capture higher-order motion patterns of groups of body joints. We split action sequences into temporal blocks, Higher-order Transformer (HoT) produces embeddings of each temporal block based on (i) the body joints, (ii) pairwise links of body joints and (iii) higher-order hyper-edges of skeleton body joints. We combine such HoT embeddings of hyper-edges of orders 1, ..., r by a novel Multi-order Multi-mode Transformer (3Mformer) with two modules whose order can be exchanged to achieve coupled

-mode attention on coupled-mode tokens based on 'channel-temporal block', 'order-channel-body joint', 'channel-hyper-edge (any order)' and 'channel-only' pairs. The first module, called Multi-order Pooling (MP), additionally learns weighted aggregation along the hyper-edge mode, whereas the second module, Temporal block Pooling (TP), aggregates along the temporal block mode. Our end-to-end trainable network yields state-of-the-art results compared to GCN-, transformer- and hypergraph-based counterparts.
****
HumanBench: Towards General Human-Centric Perception With Projector Assisted Pretraining
Shixiang Tang, Cheng Chen, Qingsong Xie, Meilin Chen, Yizhou Wang, Yuanzheng Ci, Lei Bai, Feng Zhu, Haiyang Yang, Li Yi, Rui Zhao, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21970-21982
Human-centric perceptions include a variety of vision tasks, which have widespread industrial applications, including surveillance, autonomous driving, and the metaverse. It is desirable to have a general pretrain model for versatile human-centric downstream tasks. This paper forges ahead along this path from the aspects of both benchmark and pretraining methods. Specifically, we propose a HumanBench based on existing datasets to comprehensively evaluate on the common ground the generalization abilities of different pretraining methods on 19 datasets from 6 diverse downstream tasks, including person ReID, pose estimation, human parsing, pedestrian attribute recognition, pedestrian detection, and crowd counting. To learn both coarse-grained and fine-grained knowledge in human bodies, we further propose a Projector AssisTed Hierarchical pretraining method (PATH) to learn diverse knowledge at different granularity levels. Comprehensive evaluations on HumanBench show that our PATH achieves new state-of-the-art results on 17 downstream datasets and on-par results on the other 2 datasets. The code will be publicly at https://github.com/OpenGVLab/HumanBench.
****
Heterogeneous Continual Learning
Divyam Madaan, Hongxu Yin, Wonmin Byeon, Jan Kautz, Pavlo Molchanov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15985-15995
We propose a novel framework and a solution to tackle the continual learning (CL) problem with changing network architectures. Most CL methods focus on adapting a single architecture to a new task/class by modifying its weights. However, with rapid progress in architecture design, the problem of adapting existing solutions to novel architectures becomes relevant. To address this limitation, we propose Heterogeneous Continual Learning (HCL), where a wide range of evolving network architectures emerge continually together with novel data/tasks. As a solution, we build on top of the distillation family of techniques and modify it to a new setting where a weaker model takes the role of a teacher; meanwhile, a new stronger architecture acts as a student. Furthermore, we consider a setup of limited access to previous data and propose Quick Deep Inversion (QDI) to recover prior task visual features to support knowledge transfer. QDI significantly reduces computational costs compared to previous solutions and improves overall performance. In summary, we propose a new setup for CL with a modified knowledge distillation paradigm and design a quick data inversion method to enhance distillation. Our evaluation of various benchmarks shows a significant improvement on accuracy in comparison to state-of-the-art methods over various networks architectures.
****
Object Pose Estimation With Statistical Guarantees: Conformal Keypoint Detection and Geometric Uncertainty Propagation
Heng Yang, Marco Pavone; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8947-8958
The two-stage object pose estimation paradigm first detects semantic keypoints on the image and then estimates the 6D pose by minimizing reprojection errors. Despite performing well on standard benchmarks, existing techniques offer no prova

ble guarantees on the quality and uncertainty of the estimation. In this paper, we inject two fundamental changes, namely conformal keypoint detection and geome tric uncertainty propagation, into the two-stage paradigm and propose the first pose estimator that endows an estimation with provable and computable worst-case error bounds. On one hand, conformal keypoint detection applies the statistical machinery of inductive conformal prediction to convert heuristic keypoint detec tions into circular or elliptical prediction sets that cover the groundtruth key points with a user-specified marginal probability (e.g., 90%). Geometric uncerta inty propagation, on the other, propagates the geometric constraints on the keyp oints to the 6D object pose, leading to a Pose UnceRtainty SEt (PURSE) that guar antees coverage of the groundtruth pose with the same probability. The PURSE, ho wever, is a nonconvex set that does not directly lead to estimated poses and unc ertainties. Therefore, we develop RANdom SAmple averaGing (RANSAG) to compute an average pose and apply semidefinite relaxation to upper bound the worst-case er rors between the average pose and the groundtruth. On the LineMOD Occlusion data set we demonstrate: (i) the PURSE covers the groundtruth with valid probabilitie s; (ii) the worst-case error bounds provide correct uncertainty quantification; and (iii) the average pose achieves better or similar accuracy as representative methods based on sparse keypoints.
****

Transformer Scale Gate for Semantic Segmentation

Hengcan Shi, Munawar Hayat, Jianfei Cai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3051-3060

Effectively encoding multi-scale contextual information is crucial for accurate semantic segmentation. Most of the existing transformer-based segmentation model s combine features across scales without any selection, where features on sub-op timal scales may degrade segmentation outcomes. Leveraging from the inherent pro perties of Vision Transformers, we propose a simple yet effective module, Transf ormer Scale Gate (TSG), to optimally combine multi-scale features. TSG exploits cues in self and cross attentions in Vision Transformers for the scale selection . TSG is a highly flexible plug-and-play module, and can easily be incorporated with any encoder-decoder-based hierarchical vision Transformer architecture. Ext ensive experiments on the Pascal Context, ADE20K and Cityscapes datasets demonst rate that our feature selection strategy achieves consistent gains.
****

Deep Graph Reprogramming

Yongcheng Jing, Chongbin Yuan, Li Ju, Yiding Yang, Xinchao Wang, Dacheng Tao; Pr oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24345-24354

In this paper, we explore a novel model reusing task tailored for graph neural n etworks (GNNs), termed as "deep graph reprogramming". We strive to reprogram a p re-trained GNN, without amending raw node features nor model parameters, to hand le a bunch of cross-level downstream tasks in various domains. To this end, we p ropose an innovative Data Reprogramming paradigm alongside a Model Reprogramming paradigm. The former one aims to address the challenge of diversified graph fea ture dimensions for various tasks on the input side, while the latter alleviates the dilemma of fixed per-task-per-model behavior on the model side. For data re programming, we specifically devise an elaborated Meta-FeatPadding method to dea l with heterogeneous input dimensions, and also develop a transductive Edge-Slim ming as well as an inductive Meta-GraPadding approach for diverse homogenous sam ples. Meanwhile, for model reprogramming, we propose a novel task-adaptive Repro grammable-Aggregator, to endow the frozen model with larger expressive capacitie s in handling cross-domain tasks. Experiments on fourteen datasets across node/g raph classification/regression, 3D object recognition, and distributed action re cognition, demonstrate that the proposed methods yield gratifying results, on pa r with those by re-training from scratch.
****

Compacting Binary Neural Networks by Sparse Kernel Selection

Yikai Wang, Wenbing Huang, Yinpeng Dong, Fuchun Sun, Anbang Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,

pp. 24374-24383
Binary Neural Network (BNN) represents convolution weights with 1-bit values, which enhances the efficiency of storage and computation. This paper is motivated by a previously revealed phenomenon that the binary kernels in successful BNNs are nearly power-law distributed: their values are mostly clustered into a small number of codewords. This phenomenon encourages us to compact typical BNNs and obtain further close performance through learning non-repetitive kernels within a binary kernel subspace. Specifically, we regard the binarization process as kernel grouping in terms of a binary codebook, and our task lies in learning to select a smaller subset of codewords from the full codebook. We then leverage the Gumbel-Sinkhorn technique to approximate the codeword selection process, and develop the Permutation Straight-Through Estimator (PSTE) that is able to not only optimize the selection process end-to-end but also maintain the non-repetitive occupancy of selected codewords. Experiments verify that our method reduces both the model size and bit-wise computational costs, and achieves accuracy improvements compared with state-of-the-art BNNs under comparable budgets.
****
EMT-NAS:Transferring Architectural Knowledge Between Tasks From Different Datasets
Peng Liao, Yaochu Jin, Wenli Du; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3643-3653
The success of multi-task learning (MTL) can largely be attributed to the shared representation of related tasks, allowing the models to better generalise. In deep learning, this is usually achieved by sharing a common neural network architecture and jointly training the weights. However, the joint training of weighting parameters on multiple related tasks may lead to performance degradation, known as negative transfer. To address this issue, this work proposes an evolutionary multi-tasking neural architecture search (EMT-NAS) algorithm to accelerate the search process by transferring architectural knowledge across multiple related tasks. In EMT-NAS, unlike the traditional MTL, the model for each task has a personalised network architecture and its own weights, thus offering the capability of effectively alleviating negative transfer. A fitness re-evaluation method is suggested to alleviate fluctuations in performance evaluations resulting from parameter sharing and the mini-batch gradient descent training method, thereby avoiding losing promising solutions during the search process. To rigorously verify the performance of EMT-NAS, the classification tasks used in the empirical assessments are derived from different datasets, including the CIFAR-10 and CIFAR-100, and four MedMNIST datasets. Extensive comparative experiments on different numbers of tasks demonstrate that EMT-NAS takes 8% and up to 40% on CIFAR and MedMNIST, respectively, less time to find competitive neural architectures than its single-task counterparts.
****
3D-Aware Multi-Class Image-to-Image Translation With NeRFs
Senmao Li, Joost van de Weijer, Yaxing Wang, Fahad Shahbaz Khan, Meiqin Liu, Jian Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12652-12662
Recent advances in 3D-aware generative models (3D-aware GANs) combined with Neural Radiance Fields (NeRF) have achieved impressive results. However no prior works investigate 3D-aware GANs for 3D consistent multi-class image-to-image (3D-aware I2I) translation. Naively using 2D-I2I translation methods suffers from unrealistic shape/identity change. To perform 3D-aware multi-class I2I translation, we decouple this learning process into a multi-class 3D-aware GAN step and a 3D-aware I2I translation step. In the first step, we propose two novel techniques: a new conditional architecture and an effective training strategy. In the second step, based on the well-trained multi-class 3D-aware GAN architecture, that preserves view-consistency, we construct a 3D-aware I2I translation system. To further reduce the view-consistency problems, we propose several new techniques, including a U-net-like adaptor network design, a hierarchical representation constrain and a relative regularization loss. In extensive experiments on two datasets, quantitative and qualitative results demonstrate that we successfully perform

3D-aware I2I translation with multi-view consistency.
****

Learning Joint Latent Space EBM Prior Model for Multi-Layer Generator
Jiali Cui, Ying Nian Wu, Tian Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3603-3612
This paper studies the fundamental problem of learning multi-layer generator models. The multi-layer generator model builds multiple layers of latent variables as a prior model on top of the generator, which benefits learning complex data distribution and hierarchical representations. However, such a prior model usually focuses on modeling inter-layer relations between latent variables by assuming non-informative (conditional) Gaussian distributions, which can be limited in model expressivity. To tackle this issue and learn more expressive prior models, we propose an energy-based model (EBM) on the joint latent space over all layers of latent variables with the multi-layer generator as its backbone. Such joint latent space EBM prior model captures the intra-layer contextual relations at each layer through layer-wise energy terms, and latent variables across different layers are jointly corrected. We develop a joint training scheme via maximum likelihood estimation (MLE), which involves Markov Chain Monte Carlo (MCMC) sampling for both prior and posterior distributions of the latent variables from different layers. To ensure efficient inference and learning, we further propose a variational training scheme where an inference model is used to amortize the costly posterior MCMC sampling. Our experiments demonstrate that the learned model can be expressive in generating high-quality images and capturing hierarchical features for better outlier detection.
****

Unsupervised Visible-Infrared Person Re-Identification via Progressive Graph Matching and Alternate Learning
Zesen Wu, Mang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9548-9558
Unsupervised visible-infrared person re-identification is a challenging task due to the large modality gap and the unavailability of cross-modality correspondences. Cross-modality correspondences are very crucial to bridge the modality gap. Some existing works try to mine cross-modality correspondences, but they focus only on local information. They do not fully exploit the global relationship across identities, thus limiting the quality of the mined correspondences. Worse still, the number of clusters of the two modalities is often inconsistent, exacerbating the unreliability of the generated correspondences. In response, we devise a Progressive Graph Matching method to globally mine cross-modality correspondences under cluster imbalance scenarios. PGM formulates correspondences mining as a graph matching process and considers the global information by minimizing the global matching cost, where the matching cost measures the dissimilarity of clusters. Besides, PGM adopts a progressive strategy to address the imbalance issue with multiple dynamic matching processes. Based on PGM, we design an Alternate Cross Contrastive Learning (ACCL) module to reduce the modality gap with the mined cross-modality correspondences, while mitigating the effect of noise in correspondences through an alternate scheme. Extensive experiments demonstrate the reliability of the generated correspondences and the effectiveness of our method.
****

Hierarchical B-Frame Video Coding Using Two-Layer CANF Without Motion Coding
David Alexandre, Hsueh-Ming Hang, Wen-Hsiao Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10249-10258
Typical video compression systems consist of two main modules: motion coding and residual coding. This general architecture is adopted by classical coding schemes (such as international standards H.265 and H.266) and deep learning-based coding schemes. We propose a novel B-frame coding architecture based on two-layer Conditional Augmented Normalization Flows (CANF). It has the striking feature of not transmitting any motion information. Our proposed idea of video compression without motion coding offers a new direction for learned video coding. Our base layer is a low-resolution image compressor that replaces the full-resolution mot

ion compressor. The low-resolution coded image is merged with the warped high-re solution images to generate a high-quality image as a conditioning signal for th e enhancement-layer image coding in full resolution. One advantage of this archi tecture is significantly reduced computational complexity due to eliminating the motion information compressor. In addition, we adopt a skip-mode coding techniq ue to reduce the transmitted latent samples. The rate-distortion performance of our scheme is slightly lower than that of the state-of-the-art learned B-frame c oding scheme, B-CANF, but outperforms other learned B-frame coding schemes. Howe ver, compared to B-CANF, our scheme saves 45% of multiply-accumulate operations (MACs) for encoding and 27% of MACs for decoding. The code is available at https ://nycu-clab.github.io.
****
Benchmarking Robustness of 3D Object Detection to Common Corruptions
Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, Jun Zhu; Proceedings of the IEEE/CVF Conference on Computer V ision and Pattern Recognition (CVPR), 2023, pp. 1022-1032
3D object detection is an important task in autonomous driving to perceive the s urroundings. Despite the excellent performance, the existing 3D detectors lack t he robustness to real-world corruptions caused by adverse weathers, sensor noise s, etc., provoking concerns about the safety and reliability of autonomous drivi ng systems. To comprehensively and rigorously benchmark the corruption robustnes s of 3D detectors, in this paper we design 27 types of common corruptions for bo th LiDAR and camera inputs considering real-world driving scenarios. By synthesi zing these corruptions on public datasets, we establish three corruption robustn ess benchmarks---KITTI-C, nuScenes-C, and Waymo-C. Then, we conduct large-scale experiments on 24 diverse 3D object detection models to evaluate their corruptio n robustness. Based on the evaluation results, we draw several important finding s, including: 1) motion-level corruptions are the most threatening ones that lea d to significant performance drop of all models; 2) LiDAR-camera fusion models d emonstrate better robustness; 3) camera-only models are extremely vulnerable to image corruptions, showing the indispensability of LiDAR point clouds. We releas e the benchmarks and codes at https://github.com/thu-ml/3D_Corruptions_AD to be helpful for future studies.
****
Unified Mask Embedding and Correspondence Learning for Self-Supervised Video Seg mentation
Liulei Li, Wenguan Wang, Tianfei Zhou, Jianwu Li, Yi Yang; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1 8706-18716
The objective of this paper is self-supervised learning of video object segmenta tion. We develop a unified framework which simultaneously models cross-frame den se correspondence for locally discriminative feature learning and embeds object- level context for target-mask decoding. As a result, it is able to directly lear n to perform mask-guided sequential segmentation from unlabeled videos, in contr ast to previous efforts usually relying on an oblique solution --- cheaply "copy ing" labels according to pixel-wise correlations. Concretely, our algorithm alte rnates between i) clustering video pixels for creating pseudo segmentation label s ex nihilo; and ii) utilizing the pseudo labels to learn mask encoding and deco ding for VOS. Unsupervised correspondence learning is further incorporated into this self-taught, mask embedding scheme, so as to ensure the generic nature of t he learnt representation and avoid cluster degeneracy. Our algorithm sets state- of-the-arts on two standard benchmarks (i.e., DAVIS17 and YouTube-VOS), narrowin g the gap between self- and fully-supervised VOS, in terms of both performance a nd network architecture design. Our full code will be released.
****
Seeing Beyond the Brain: Conditional Diffusion Model With Sparse Masked Modeling for Vision Decoding
Zijiao Chen, Jiaxin Qing, Tiange Xiang, Wan Lin Yue, Juan Helen Zhou; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22710-22720

Decoding visual stimuli from brain recordings aims to deepen our understanding o
f the human visual system and build a solid foundation for bridging human and co
mputer vision through the Brain-Computer Interface. However, reconstructing high
-quality images with correct semantics from brain recordings is a challenging pr
oblem due to the complex underlying representations of brain signals and the sca
rcity of data annotations. In this work, we present MinD-Vis: Sparse Masked Brai
n Modeling with Double-Conditioned Latent Diffusion Model for Human Vision Decod
ing. Firstly, we learn an effective self-supervised representation of fMRI data
using mask modeling in a large latent space inspired by the sparse coding of inf
ormation in the primary visual cortex. Then by augmenting a latent diffusion mod
el with double-conditioning, we show that MinD-Vis can reconstruct highly plausi
ble images with semantically matching details from brain recordings using very f
ew paired annotations. We benchmarked our model qualitatively and quantitatively
; the experimental results indicate that our method outperformed state-of-the-ar
t in both semantic mapping (100-way semantic classification) and generation qual
ity (FID) by 66% and 41% respectively. An exhaustive ablation study was also con
ducted to analyze our framework.
****
PointAvatar: Deformable Point-Based Head Avatars From Videos
Yufeng Zheng, Wang Yifan, Gordon Wetzstein, Michael J. Black, Otmar Hilliges; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2023, pp. 21057-21067
The ability to create realistic animatable and relightable head avatars from cas
ual video sequences would open up wide ranging applications in communication and
 entertainment. Current methods either build on explicit 3D morphable meshes (3D
MM) or exploit neural implicit representations. The former are limited by fixed
topology, while the latter are non-trivial to deform and inefficient to render.
Furthermore, existing approaches entangle lighting and albedo, limiting the abil
ity to re-render the avatar in new environments. In contrast, we propose PointAv
atar, a deformable point-based representation that disentangles the source color
 into intrinsic albedo and normal-dependent shading. We demonstrate that PointAv
atar bridges the gap between existing mesh- and implicit representations, combin
ing high-quality geometry and appearance with topological flexibility, ease of d
eformation and rendering efficiency. We show that our method is able to generate
 animatable 3D avatars using monocular videos from multiple sources including ha
nd-held smartphones, laptop webcams and internet videos, achieving state-of-the-
art quality in challenging cases where previous methods fail, e.g., thin hair st
rands, while being significantly more efficient in training than competing metho
ds.
****
Seeing Through the Glass: Neural 3D Reconstruction of Object Inside a Transparen
t Container
Jinguang Tong, Sundaram Muthu, Fahira Afzal Maken, Chuong Nguyen, Hongdong Li; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2023, pp. 12555-12564
In this paper, we define a new problem of recovering the 3D geometry of an objec
t confined in a transparent enclosure. We also propose a novel method for solvin
g this challenging problem. Transparent enclosures pose challenges of multiple l
ight reflections and refractions at the interface between different propagation
media e.g. air or glass. These multiple reflections and refractions cause seriou
s image distortions which invalidate the single viewpoint assumption. Hence the
3D geometry of such objects cannot be reliably reconstructed using existing meth
ods, such as traditional structure from motion or modern neural reconstruction m
ethods. We solve this problem by explicitly modeling the scene as two distinct s
ub-spaces, inside and outside the transparent enclosure. We use an existing neur
al reconstruction method (NeuS) that implicitly represents the geometry and appe
arance of the inner subspace. In order to account for complex light interactions
, we develop a hybrid rendering strategy that combines volume rendering with ray
 tracing. We then recover the underlying geometry and appearance of the model by
 minimizing the difference between the real and rendered images. We evaluate our

method on both synthetic and real data. Experiment results show that our method outperforms the state-of-the-art (SOTA) methods. Codes and data will be available at https://github.com/hirotong/ReNeuS
****
OrienterNet: Visual Localization in 2D Public Maps With Neural Matching
Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, Vasileios Balntas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21632-21642
Humans can orient themselves in their 3D environments using simple 2D maps. Differently, algorithms for visual localization mostly rely on complex 3D point clouds that are expensive to build, store, and maintain over time. We bridge this gap by introducing OrienterNet, the first deep neural network that can localize an image with sub-meter accuracy using the same 2D semantic maps that humans use. OrienterNet estimates the location and orientation of a query image by matching a neural Bird's-Eye View with open and globally available maps from OpenStreetMap, enabling anyone to localize anywhere such maps are available. OrienterNet is supervised only by camera poses but learns to perform semantic matching with a wide range of map elements in an end-to-end manner. To enable this, we introduce a large crowd-sourced dataset of images captured across 12 cities from the diverse viewpoints of cars, bikes, and pedestrians. OrienterNet generalizes to new datasets and pushes the state of the art in both robotics and AR scenarios. The code is available at https://github.com/facebookresearch/OrienterNet
****
PMatch: Paired Masked Image Modeling for Dense Geometric Matching
Shengjie Zhu, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21909-21918
Dense geometric matching determines the dense pixel-wise correspondence between a source and support image corresponding to the same 3D structure. Prior works employ an encoder of transformer blocks to correlate the two-frame features. However, existing monocular pretraining tasks, e.g., image classification, and masked image modeling (MIM), can not pretrain the cross-frame module, yielding less optimal performance. To resolve this, we reformulate the MIM from reconstructing a single masked image to reconstructing a pair of masked images, enabling the pretraining of transformer module. Additionally, we incorporate a decoder into pretraining for improved upsampling results. Further, to be robust to the textureless area, we propose a novel cross-frame global matching module (CFGM). Since the most textureless area is planar surfaces, we propose a homography loss to further regularize its learning. Combined together, we achieve the State-of-The-Art (SoTA) performance on geometric matching. Codes and models are available at https://github.com/ShngJZ/PMatch.
****
Neural Voting Field for Camera-Space 3D Hand Pose Estimation
Lin Huang, Chung-Ching Lin, Kevin Lin, Lin Liang, Lijuan Wang, Junsong Yuan, Zicheng Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8969-8978
We present a unified framework for camera-space 3D hand pose estimation from a single RGB image based on 3D implicit representation. As opposed to recent works, most of which first adopt holistic or pixel-level dense regression to obtain relative 3D hand pose and then follow with complex second-stage operations for 3D global root or scale recovery, we propose a novel unified 3D dense regression scheme to estimate camera-space 3D hand pose via dense 3D point-wise voting in camera frustum. Through direct dense modeling in 3D domain inspired by Pixel-aligned Implicit Functions for 3D detailed reconstruction, our proposed Neural Voting Field (NVF) fully models 3D dense local evidence and hand global geometry, helping to alleviate common 2D-to-3D ambiguities. Specifically, for a 3D query point in camera frustum and its pixel-aligned image feature, NVF, represented by a Multi-Layer Perceptron, regresses: (i) its signed distance to the hand surface; (ii) a set of 4D offset vectors (1D voting weight and 3D directional vector to each hand joint). Following a vote-casting scheme, 4D offset vectors from near-surfa

ce points are selected to calculate the 3D hand joint coordinates by a weighted average. Experiments demonstrate that NVF outperforms existing state-of-the-art algorithms on FreiHAND dataset for camera-space 3D hand pose estimation. We also adapt NVF to the classic task of root-relative 3D hand pose estimation, for which NVF also obtains state-of-the-art results on HO3D dataset.
****
STMT: A Spatial-Temporal Mesh Transformer for MoCap-Based Action Recognition
Xiaoyu Zhu, Po-Yao Huang, Junwei Liang, Celso M. de Melo, Alexander G. Hauptmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1526-1536
We study the problem of human action recognition using motion capture (MoCap) sequences. Unlike existing techniques that take multiple manual steps to derive standardized skeleton representations as model input, we propose a novel Spatial-Temporal Mesh Transformer (STMT) to directly model the mesh sequences. The model uses a hierarchical transformer with intra-frame off-set attention and inter-frame self-attention. The attention mechanism allows the model to freely attend between any two vertex patches to learn non-local relationships in the spatial-temporal domain. Masked vertex modeling and future frame prediction are used as two self-supervised tasks to fully activate the bi-directional and auto-regressive attention in our hierarchical transformer. The proposed method achieves state-of-the-art performance compared to skeleton-based and point-cloud-based models on common MoCap benchmarks. Code is available at https://github.com/zgzxy001/STMT.
****
Visual Recognition-Driven Image Restoration for Multiple Degradation With Intrinsic Semantics Recovery
Zizheng Yang, Jie Huang, Jiahao Chang, Man Zhou, Hu Yu, Jinghao Zhang, Feng Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14059-14070
Deep image recognition models suffer a significant performance drop when applied to low-quality images since they are trained on high-quality images. Although many studies have investigated to solve the issue through image restoration or domain adaptation, the former focuses on visual quality rather than recognition quality, while the latter requires semantic annotations for task-specific training. In this paper, to address more practical scenarios, we propose a Visual Recognition-Driven Image Restoration network for multiple degradation, dubbed VRD-IR, to recover high-quality images from various unknown corruption types from the perspective of visual recognition within one model. Concretely, we harmonize the semantic representations of diverse degraded images into a unified space in a dynamic manner, and then optimize them towards intrinsic semantics recovery. Moreover, a prior-ascribing optimization strategy is introduced to encourage VRD-IR to couple with various downstream recognition tasks better. Our VRD-IR is corruption- and recognition-agnostic, and can be inserted into various recognition tasks directly as an image enhancement module. Extensive experiments on multiple image distortions demonstrate that our VRD-IR surpasses existing image restoration methods and show superior performance on diverse high-level tasks, including classification, detection, and person re-identification.
****
High-Fidelity Generalized Emotional Talking Face Generation With Multi-Modal Emotion Space Learning
Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, Yong Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6609-6619
Recently, emotional talking face generation has received considerable attention. However, existing methods only adopt one-hot coding, image, or audio as emotion conditions, thus lacking flexible control in practical applications and failing to handle unseen emotion styles due to limited semantics. They either ignore the one-shot setting or the quality of generated faces. In this paper, we propose a more flexible and generalized framework. Specifically, we supplement the emotion style in text prompts and use an Aligned Multi-modal Emotion encoder to embed the text, image, and audio emotion modality into a unified space, which inherit

s rich semantic prior from CLIP. Consequently, effective multi-modal emotion space learning helps our method support arbitrary emotion modality during testing and could generalize to unseen emotion styles. Besides, an Emotion-aware Audio-to-3DMM Convertor is proposed to connect the emotion condition and the audio sequence to structural representation. A followed style-based High-fidelity Emotional Face generator is designed to generate arbitrary high-resolution realistic identities. Our texture generator hierarchically learns flow fields and animated faces in a residual manner. Extensive experiments demonstrate the flexibility and generalization of our method in emotion control and the effectiveness of high-quality face synthesis.
****
Masked and Adaptive Transformer for Exemplar Based Image Translation
Chang Jiang, Fei Gao, Biao Ma, Yuhao Lin, Nannan Wang, Gang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22418-22427
We present a novel framework for exemplar based image translation. Recent advanced methods for this task mainly focus on establishing cross-domain semantic correspondence, which sequentially dominates image generation in the manner of local style control. Unfortunately, cross domain semantic matching is challenging; and matching errors ultimately degrade the quality of generated images. To overcome this challenge, we improve the accuracy of matching on the one hand, and diminish the role of matching in image generation on the other hand. To achieve the former, we propose a masked and adaptive transformer (MAT) for learning accurate cross-domain correspondence, and executing context-aware feature augmentation. To achieve the latter, we use source features of the input and global style codes of the exemplar, as supplementary information, for decoding an image. Besides, we devise a novel contrastive style learning method, for acquire quality-discriminative style representations, which in turn benefit high-quality image generation. Experimental results show that our method, dubbed MATEBIT, performs considerably better than state-of-the-art methods, in diverse image translation tasks.
****
Knowledge Combination To Learn Rotated Detection Without Rotated Annotation
Tianyu Zhu, Bryce Ferenczi, Pulak Purkait, Tom Drummond, Hamid Rezatofighi, Anton van den Hengel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15518-15527
Rotated bounding boxes drastically reduce output ambiguity of elongated objects, making it superior to axis-aligned bounding boxes. Despite the effectiveness, rotated detectors are not widely employed. Annotating rotated bounding boxes is such a laborious process that they are not provided in many detection datasets where axis-aligned annotations are used instead. In this paper, we propose a framework that allows the model to predict precise rotated boxes only requiring cheaper axis-aligned annotation of the target dataset. To achieve this, we leverage the fact that neural networks are capable of learning richer representation of the target domain than what is utilized by the task. The under-utilized representation can be exploited to address a more detailed task. Our framework combines task knowledge of an out-of-domain source dataset with stronger annotation and domain knowledge of the target dataset with weaker annotation. A novel assignment process and projection loss are used to enable the co-training on the source and target datasets. As a result, the model is able to solve the more detailed task in the target domain, without additional computation overhead during inference. We extensively evaluate the method on various target datasets including fresh-produce dataset, HRSC2016 and SSDD. Results show that the proposed method consistently performs on par with the fully supervised approach.
****
Teaching Matters: Investigating the Role of Supervision in Vision Transformers
Matthew Walmer, Saksham Suri, Kamal Gupta, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7486-7496
Vision Transformers (ViTs) have gained significant popularity in recent years and have proliferated into many applications. However, their behavior under differ

ent learning paradigms is not well explored. We compare ViTs trained through dif ferent methods of supervision, and show that they learn a diverse range of behav iors in terms of their attention, representations, and downstream performance. W e also discover ViT behaviors that are consistent across supervision, including the emergence of Offset Local Attention Heads. These are self-attention heads th at attend to a token adjacent to the current token with a fixed directional offs et, a phenomenon that to the best of our knowledge has not been highlighted in a ny prior work. Our analysis shows that ViTs are highly flexible and learn to pro cess local and global information in different orders depending on their trainin g method. We find that contrastive self-supervised methods learn features that a re competitive with explicitly supervised features, and they can even be superio r for part-level tasks. We also find that the representations of reconstruction-based models show non-trivial similarity to contrastive self-supervised models.
****

Imagic: Text-Based Real Image Editing With Diffusion Models

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, Michal Irani; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6007-6017

Text-conditioned image editing has recently attracted considerable interest. How ever, most methods are currently limited to one of the following: specific editi ng types (e.g., object overlay, style transfer), synthetically generated images, or requiring multiple input images of a common object. In this paper we demonst rate, for the very first time, the ability to apply complex (e.g., non-rigid) te xt-based semantic edits to a single real image. For example, we can change the p osture and composition of one or multiple objects inside an image, while preserv ing its original characteristics. Our method can make a standing dog sit down, c ause a bird to spread its wings, etc. -- each within its single high-resolution user-provided natural image. Contrary to previous work, our proposed method requ ires only a single input image and a target text (the desired edit). It operates on real images, and does not require any additional inputs (such as image masks or additional views of the object). Our method, called Imagic, leverages a pre-trained text-to-image diffusion model for this task. It produces a text embeddin g that aligns with both the input image and the target text, while fine-tuning t he diffusion model to capture the image-specific appearance. We demonstrate the quality and versatility of Imagic on numerous inputs from various domains, showc asing a plethora of high quality complex semantic image edits, all within a sing le unified framework. To better assess performance, we introduce TEdBench, a hig hly challenging image editing benchmark. We conduct a user study, whose findings show that human raters prefer Imagic to previous leading editing methods on TEd Bench.
****

Pointersect: Neural Rendering With Cloud-Ray Intersection

Jen-Hao Rick Chang, Wei-Yu Chen, Anurag Ranjan, Kwang Moo Yi, Oncel Tuzel; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2023, pp. 8359-8369

We propose a novel method that renders point clouds as if they are surfaces. The proposed method is differentiable and requires no scene-specific optimization. This unique capability enables, out-of-the-box, surface normal estimation, rende ring room-scale point clouds, inverse rendering, and ray tracing with global ill umination. Unlike existing work that focuses on converting point clouds to other representations--e.g., surfaces or implicit functions--our key idea is to direc tly infer the intersection of a light ray with the underlying surface represente d by the given point cloud. Specifically, we train a set transformer that, given a small number of local neighbor points along a light ray, provides the interse ction point, the surface normal, and the material blending weights, which are us ed to render the outcome of this light ray. Localizing the problem into small ne ighborhoods enables us to train a model with only 48 meshes and apply it to unse en point clouds. Our model achieves higher estimation accuracy than state-of-the -art surface reconstruction and point-cloud rendering methods on three test sets . When applied to room-scale point clouds, without any scene-specific optimizati

on, the model achieves competitive quality with the state-of-the-art novel-view rendering methods. Moreover, we demonstrate ability to render and manipulate Lidar-scanned point clouds such as lighting control and object insertion.
****
Beyond Attentive Tokens: Incorporating Token Importance and Diversity for Efficient Vision Transformers
Sifan Long, Zhen Zhao, Jimin Pi, Shengsheng Wang, Jingdong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10334-10343
Vision transformers have achieved significant improvements on various vision tasks but their quadratic interactions between tokens significantly reduce computational efficiency. Many pruning methods have been proposed to remove redundant tokens for efficient vision transformers recently. However, existing studies mainly focus on the token importance to preserve local attentive tokens but completely ignore the global token diversity. In this paper, we emphasize the cruciality of diverse global semantics and propose an efficient token decoupling and merging method that can jointly consider the token importance and diversity for token pruning. According to the class token attention, we decouple the attentive and inattentive tokens. In addition to preserve the most discriminative local tokens, we merge similar inattentive tokens and match homogeneous attentive tokens to maximize the token diversity. Despite its simplicity, our method obtains a promising trade-off between model complexity and classification accuracy. On DeiT-S, our method reduces the FLOPs by 35% with only a 0.2% accuracy drop. Notably, benefiting from maintaining the token diversity, our method can even improve the accuracy of DeiT-T by 0.1% after reducing its FLOPs by 40%.
****
You Are Catching My Attention: Are Vision Transformers Bad Learners Under Backdoor Attacks?
Zenghui Yuan, Pan Zhou, Kai Zou, Yu Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24605-24615
Vision Transformers (ViTs), which made a splash in the field of computer vision (CV), have shaken the dominance of convolutional neural networks (CNNs). However, in the process of industrializing ViTs, backdoor attacks have brought severe challenges to security. The success of ViTs benefits from the self-attention mechanism. However, compared with CNNs, we find that this mechanism of capturing global information within patches makes ViTs more sensitive to patch-wise triggers. Under such observations, we delicately design a novel backdoor attack framework for ViTs, dubbed BadViT, which utilizes a universal patch-wise trigger to catch the model's attention from patches beneficial for classification to those with triggers, thereby manipulating the mechanism on which ViTs survive to confuse itself. Furthermore, we propose invisible variants of BadViT to increase the stealth of the attack by limiting the strength of the trigger perturbation. Through a large number of experiments, it is proved that BadViT is an efficient backdoor attack method against ViTs, which is less dependent on the number of poisons, with satisfactory convergence, and is transferable for downstream tasks. Furthermore, the risks inside of ViTs to backdoor attacks are also explored from the perspective of existing advanced defense schemes.
****
STDLens: Model Hijacking-Resilient Federated Learning for Object Detection
Ka-Ho Chow, Ling Liu, Wenqi Wei, Fatih Ilhan, Yanzhao Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16343-16351
Federated Learning (FL) has been gaining popularity as a collaborative learning framework to train deep learning-based object detection models over a distributed population of clients. Despite its advantages, FL is vulnerable to model hijacking. The attacker can control how the object detection system should misbehave by implanting Trojaned gradients using only a small number of compromised clients in the collaborative learning process. This paper introduces STDLens, a principled approach to safeguarding FL against such attacks. We first investigate existing mitigation mechanisms and analyze their failures caused by the inherent err

ors in spatial clustering analysis on gradients. Based on the insights, we intro duce a three-tier forensic framework to identify and expel Trojaned gradients an d reclaim the performance over the course of FL. We consider three types of adap tive attacks and demonstrate the robustness of STDLens against advanced adversar ies. Extensive experiments show that STDLens can protect FL against different mo del hijacking attacks and outperform existing methods in identifying and removin g Trojaned gradients with significantly higher precision and much lower false-po sitive rates. The source code is available at https://github.com/git-disl/STDLen s.
****
Contrastive Grouping With Transformer for Referring Image Segmentation
Jiajin Tang, Ge Zheng, Cheng Shi, Sibei Yang; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23570-23580
Referring image segmentation aims to segment the target referent in an image con ditioning on a natural language expression. Existing one-stage methods employ pe r-pixel classification frameworks, which attempt straightforwardly to align visi on and language at the pixel level, thus failing to capture critical object-leve l information. In this paper, we propose a mask classification framework, Contra stive Grouping with Transformer network (CGFormer), which explicitly captures ob ject-level information via token-based querying and grouping strategy. Specifica lly, CGFormer first introduces learnable query tokens to represent objects and t hen alternately queries linguistic features and groups visual features into the query tokens for object-aware cross-modal reasoning. In addition, CGFormer achie ves cross-level interaction by jointly updating the query tokens and decoding ma sks in every two consecutive layers. Finally, CGFormer cooperates contrastive le arning to the grouping strategy to identify the token and its mask corresponding to the referent. Experimental results demonstrate that CGFormer outperforms sta te-of-the-art methods in both segmentation and generalization settings consisten tly and significantly. Code is available at https://github.com/Toneyaya/CGFormer .
****
MagicPony: Learning Articulated 3D Animals in the Wild
Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, Andrea Vedaldi; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2023, pp. 8792-8802
We consider the problem of predicting the 3D shape, articulation, viewpoint, tex ture, and lighting of an articulated animal like a horse given a single test ima ge as input. We present a new method, dubbed MagicPony, that learns this predict or purely from in-the-wild single-view images of the object category, with minim al assumptions about the topology of deformation. At its core is an implicit-exp licit representation of articulated shape and appearance, combining the strength s of neural fields and meshes. In order to help the model understand an object's shape and pose, we distil the knowledge captured by an off-the-shelf self-super vised vision transformer and fuse it into the 3D model. To overcome local optima in viewpoint estimation, we further introduce a new viewpoint sampling scheme t hat comes at no additional training cost. MagicPony outperforms prior work on th is challenging task and demonstrates excellent generalisation in reconstructing art, despite the fact that it is only trained on real images. The code can be fo und on the project page at https://3dmagicpony.github.io/.
****
PaCa-ViT: Learning Patch-to-Cluster Attention in Vision Transformers
Ryan Grainger, Thomas Paniagua, Xi Song, Naresh Cuntoor, Mun Wai Lee, Tianfu Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2023, pp. 18568-18578
Vision Transformers (ViTs) are built on the assumption of treating image patches as "visual tokens" and learn patch-to-patch attention. The patch embedding base d tokenizer has a semantic gap with respect to its counterpart, the textual toke nizer. The patch-to-patch attention suffers from the quadratic complexity issue, and also makes it non-trivial to explain learned ViTs. To address these issues in ViT, this paper proposes to learn Patch-to-Cluster attention (PaCa) in ViT. Q

ueries in our PaCa-ViT starts with patches, while keys and values are directly based on clustering (with a predefined small number of clusters). The clusters are learned end-to-end, leading to better tokenizers and inducing joint clustering-for-attention and attention-for-clustering for better and interpretable models. The quadratic complexity is relaxed to linear complexity. The proposed PaCa module is used in designing efficient and interpretable ViT backbones and semantic segmentation head networks. In experiments, the proposed methods are tested on ImageNet-1k image classification, MS-COCO object detection and instance segmentation and MIT-ADE20k semantic segmentation. Compared with the prior art, it obtains better performance in all the three benchmarks than the SWin and the PVTs by significant margins in ImageNet-1k and MIT-ADE20k. It is also significantly more efficient than PVT models in MS-COCO and MIT-ADE20k due to the linear complexity. The learned clusters are semantically meaningful. Code and model checkpoints are available at https://github.com/iVMCL/PaCaViT.
****
Pix2map: Cross-Modal Retrieval for Inferring Street Maps From Images
Xindi Wu, KwunFung Lau, Francesco Ferroni, Aljoša Ošep, Deva Ramanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17514-17523
Self-driving vehicles rely on urban street maps for autonomous navigation. In this paper, we introduce Pix2Map, a method for inferring urban street map topology directly from ego-view images, as needed to continually update and expand existing maps. This is a challenging task, as we need to infer a complex urban road topology directly from raw image data. The main insight of this paper is that this problem can be posed as cross-modal retrieval by learning a joint, cross-modal embedding space for images and existing maps, represented as discrete graphs that encode the topological layout of the visual surroundings. We conduct our experimental evaluation using the Argoverse dataset and show that it is indeed possible to accurately retrieve street maps corresponding to both seen and unseen roads solely from image data. Moreover, we show that our retrieved maps can be used to update or expand existing maps and even show proof-of-concept results for visual localization and image retrieval from spatial graphs.
****
LightPainter: Interactive Portrait Relighting With Freehand Scribble
Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 195-205
Recent portrait relighting methods have achieved realistic results of portrait lighting effects given a desired lighting representation such as an environment map. However, these methods are not intuitive for user interaction and lack precise lighting control. We introduce LightPainter, a scribble-based relighting system that allows users to interactively manipulate portrait lighting effect with ease. This is achieved by two conditional neural networks, a delighting module that recovers geometry and albedo optionally conditioned on skin tone, and a scribble-based module for relighting. To train the relighting module, we propose a novel scribble simulation procedure to mimic real user scribbles, which allows our pipeline to be trained without any human annotations. We demonstrate high-quality and flexible portrait lighting editing capability with both quantitative and qualitative experiments. User study comparisons with commercial lighting editing tools also demonstrate consistent user preference for our method.
****
Affordances From Human Videos as a Versatile Representation for Robotics
Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, Deepak Pathak; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13778-13790
Building a robot that can understand and learn to interact by watching humans has inspired several vision problems. However, despite some successful results on static datasets, it remains unclear how current models can be used on a robot directly. In this paper, we aim to bridge this gap by leveraging videos of human interactions in an environment centric manner. Utilizing internet videos of human

behavior, we train a visual affordance model that estimates where and how in the scene a human is likely to interact. The structure of these behavioral affordances directly enables the robot to perform many complex tasks. We show how to seamlessly integrate our affordance model with four robot learning paradigms including offline imitation learning, exploration, goal-conditioned learning, and action parameterization for reinforcement learning. We show the efficacy of our approach, which we call Vision-Robotics Bridge (VRB) across 4 real world environments, over 10 different tasks, and 2 robotic platforms operating in the wild.
****
Unsupervised Inference of Signed Distance Functions From Single Sparse Point Clouds Without Learning Priors
Chao Chen, Yu-Shen Liu, Zhizhong Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17712-17723
It is vital to infer signed distance functions (SDFs) from 3D point clouds. The latest methods rely on generalizing the priors learned from large scale supervision. However, the learned priors do not generalize well to various geometric variations that are unseen during training, especially for extremely sparse point clouds. To resolve this issue, we present a neural network to directly infer SDFs from single sparse point clouds without using signed distance supervision, learned priors or even normals. Our insight here is to learn surface parameterization and SDFs inference in an end-to-end manner. To make up the sparsity, we leverage parameterized surfaces as a coarse surface sampler to provide many coarse surface estimations in training iterations, according to which we mine supervision and our thin plate splines (TPS) based network infers SDFs as smooth functions in a statistical way. Our method significantly improves the generalization ability and accuracy in unseen point clouds. Our experimental results show our advantages over the state-of-the-art methods in surface reconstruction for sparse point clouds under synthetic datasets and real scans.The code is available at https://github.com/chenchao15/NeuralTPS.
****
AMT: All-Pairs Multi-Field Transforms for Efficient Frame Interpolation
Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, Ming-Ming Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9801-9810
We present All-Pairs Multi-Field Transforms (AMT), a new network architecture for video frame interpolation. It is based on two essential designs. First, we build bidirectional correlation volumes for all pairs of pixels and use the predicted bilateral flows to retrieve correlations for updating both flows and the interpolated content feature. Second, we derive multiple groups of fine-grained flow fields from one pair of updated coarse flows for performing backward warping on the input frames separately. Combining these two designs enables us to generate promising task-oriented flows and reduce the difficulties in modeling large motions and handling occluded areas during frame interpolation. These qualities promote our model to achieve state-of-the-art performance on various benchmarks with high efficiency. Moreover, our convolution-based model competes favorably compared to Transformer-based models in terms of accuracy and efficiency. Our code is available at https://github.com/MCG-NKU/AMT.
****
Vision Transformers Are Parameter-Efficient Audio-Visual Learners
Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, Gedas Bertasius; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2299-2309
Vision transformers (ViTs) have achieved impressive results on various computer vision tasks in the last several years. In this work, we study the capability of frozen ViTs, pretrained only on visual data, to generalize to audio-visual data without finetuning any of its original parameters. To do so, we propose a latent audio-visual hybrid (LAVISH) adapter that adapts pretrained ViTs to audio-visual tasks by injecting a small number of trainable parameters into every layer of a frozen ViT. To efficiently fuse visual and audio cues, our LAVISH adapter uses a small set of latent tokens, which form an attention bottleneck, thus, elimin

ating the quadratic cost of standard cross-attention. Compared to the existing modality-specific audio-visual methods, our approach achieves competitive or even better performance on various audio-visual tasks while using fewer tunable parameters and without relying on costly audio pretraining or external audio encoders. Our code is available at https://genjib.github.io/project_page/LAVISH/
****
Deep Discriminative Spatial and Temporal Network for Efficient Video Deblurring
Jinshan Pan, Boming Xu, Jiangxin Dong, Jianjun Ge, Jinhui Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22191-22200
How to effectively explore spatial and temporal information is important for video deblurring. In contrast to existing methods that directly align adjacent frames without discrimination, we develop a deep discriminative spatial and temporal network to facilitate the spatial and temporal feature exploration for better video deblurring. We first develop a channel-wise gated dynamic network to adaptively explore the spatial information. As adjacent frames usually contain different contents, directly stacking features of adjacent frames without discrimination may affect the latent clear frame restoration. Therefore, we develop a simple yet effective discriminative temporal feature fusion module to obtain useful temporal features for latent frame restoration. Moreover, to utilize the information from long-range frames, we develop a wavelet-based feature propagation method that takes the discriminative temporal feature fusion module as the basic unit to effectively propagate main structures from long-range frames for better video deblurring. We show that the proposed method does not require additional alignment methods and performs favorably against state-of-the-art ones on benchmark datasets in terms of accuracy and model complexity.
****
Training Debiased Subnetworks With Contrastive Weight Pruning
Geon Yeong Park, Sangmin Lee, Sang Wan Lee, Jong Chul Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7929-7938
Neural networks are often biased to spuriously correlated features that provide misleading statistical evidence that does not generalize. This raises an interesting question: "Does an optimal unbiased functional subnetwork exist in a severely biased network? If so, how to extract such subnetwork?" While empirical evidence has been accumulated about the existence of such unbiased subnetworks, these observations are mainly based on the guidance of ground-truth unbiased samples. Thus, it is unexplored how to discover the optimal subnetworks with biased training datasets in practice. To address this, here we first present our theoretical insight that alerts potential limitations of existing algorithms in exploring unbiased subnetworks in the presence of strong spurious correlations. We then further elucidate the importance of bias-conflicting samples on structure learning. Motivated by these observations, we propose a Debiased Contrastive Weight Pruning (DCWP) algorithm, which probes unbiased subnetworks without expensive group annotations. Experimental results demonstrate that our approach significantly outperforms state-of-the-art debiasing methods despite its considerable reduction in the number of parameters.
****
SparseViT: Revisiting Activation Sparsity for Efficient High-Resolution Vision Transformer
Xuanyao Chen, Zhijian Liu, Haotian Tang, Li Yi, Hang Zhao, Song Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2061-2070
High-resolution images enable neural networks to learn richer visual representations. However, this improved performance comes at the cost of growing computational complexity, hindering their usage in latency-sensitive applications. As not all pixels are equal, skipping computations for less-important regions offers a simple and effective measure to reduce the computation. This, however, is hard to be translated into actual speedup for CNNs since it breaks the regularity of the dense convolution workload. In this paper, we introduce SparseViT that revisi

ts activation sparsity for recent window-based vision transformers (ViTs). As wi
ndow attentions are naturally batched over blocks, actual speedup with window ac
tivation pruning becomes possible: i.e., 50% latency reduction with 60% sparsit
y. Different layers should be assigned with different pruning ratios due to thei
r diverse sensitivities and computational costs. We introduce sparsity-aware ada
ptation and apply the evolutionary search to efficiently find the optimal layerw
ise sparsity configuration within the vast search space. SparseViT achieves spee
dups of 1.5x, 1.4x, and 1.3x compared to its dense counterpart in monocular 3D o
bject detection, 2D instance segmentation, and 2D semantic segmentation, respect
ively, with negligible to no loss of accuracy.
****
Prototype-Based Embedding Network for Scene Graph Generation
Chaofan Zheng, Xinyu Lyu, Lianli Gao, Bo Dai, Jingkuan Song; Proceedings of the
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
 22783-22792
Current Scene Graph Generation (SGG) methods explore contextual information to p
redict relationships among entity pairs. However, due to the diverse visual appe
arance of numerous possible subject-object combinations, there is a large intra-
class variation within each predicate category, e.g., "man-eating-pizza, giraffe
-eating-leaf", and the severe inter-class similarity between different classes,
e.g., "man-holding-plate, man-eating-pizza", in model's latent space. The above
challenges prevent current SGG methods from acquiring robust features for reliab
le relation prediction. In this paper, we claim that predicate's categoryinheren
t semantics can serve as class-wise prototypes in the semantic space for relievi
ng the above challenges caused by the diverse visual appearances. To the end, we
 propose the Prototype-based Embedding Network (PE-Net), which models entities/p
redicates with prototype-aligned compact and distinctive representations and est
ablishes matching between entity pairs and predicates in a common embedding spac
e for relation recognition. Moreover, Prototypeguided Learning (PL) is introduce
d to help PE-Net efficiently learn such entity-predicate matching, and Prototype
 Regularization (PR) is devised to relieve the ambiguous entity-predicate matchi
ng caused by the predicate's semantic overlap. Extensive experiments demonstrate
 that our method gains superior relation recognition capability on SGG, achievin
g new state-of-the-art performances on both Visual Genome and Open Images datase
ts.
****
Toward RAW Object Detection: A New Benchmark and a New Model
Ruikang Xu, Chang Chen, Jingyang Peng, Cheng Li, Yibin Huang, Fenglong Song, You
liang Yan, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2023, pp. 13384-13393
In many computer vision applications (e.g., robotics and autonomous driving), hi
gh dynamic range (HDR) data is necessary for object detection algorithms to hand
le a variety of lighting conditions, such as strong glare. In this paper, we aim
 to achieve object detection on RAW sensor data, which naturally saves the HDR i
nformation from image sensors without extra equipment costs. We build a novel RA
W sensor dataset, named ROD, for Deep Neural Networks (DNNs)-based object detect
ion algorithms to be applied to HDR data. The ROD dataset contains a large amoun
t of annotated instances of day and night driving scenes in 24-bit dynamic range
. Based on the dataset, we first investigate the impact of dynamic range for DNN
s-based detectors and demonstrate the importance of dynamic range adjustment for
 detection on RAW sensor data. Then, we propose a simple and effective adjustmen
t method for object detection on HDR RAW sensor data, which is image adaptive an
d jointly optimized with the downstream detector in an end-to-end scheme. Extens
ive experiments demonstrate that the performance of detection on RAW sensor data
 is significantly superior to standard dynamic range (SDR) data in different sit
uations. Moreover, we analyze the influence of texture information and pixel dis
tribution of input data on the performance of the DNNs-based detector.
****
Music-Driven Group Choreography
Nhat Le, Thang Pham, Tuong Do, Erman Tjiputra, Quang D. Tran, Anh Nguyen; Procee

dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8673-8682
Music-driven choreography is a challenging problem with a wide variety of industrial applications. Recently, many methods have been proposed to synthesize dance motions from music for a single dancer. However, generating dance motion for a group remains an open problem. In this paper, we present AIOZ-GDANCE, a new largescale dataset for music-driven group dance generation. Unlike existing datasets that only support single dance, our new dataset contains group dance videos, hence supporting the study of group choreography. We propose a semiautonomous labeling method with humans in the loop to obtain the 3D ground truth for our dataset. The proposed dataset consists of 16.7 hours of paired music and 3D motion from in-the-wild videos, covering 7 dance styles and 16 music genres. We show that naively applying single dance generation technique to creating group dance motion may lead to unsatisfactory results, such as inconsistent movements and collisions between dancers. Based on our new dataset, we propose a new method that takes an input music sequence and a set of 3D positions of dancers to efficiently produce multiple group-coherent choreographies. We propose new evaluation metrics for measuring group dance quality and perform intensive experiments to demonstrate the effectiveness of our method. Our project facilitates future research on group dance generation and is available at https://aioz-ai.github.io/AIOZ-GDANCE/.
****
Cascade Evidential Learning for Open-World Weakly-Supervised Temporal Action Localization
Mengyuan Chen, Junyu Gao, Changsheng Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14741-14750
Targeting at recognizing and localizing action instances with only video-level labels during training, Weakly-supervised Temporal Action Localization (WTAL) has achieved significant progress in recent years. However, living in the dynamically changing open world where unknown actions constantly spring up, the closed-set assumption of existing WTAL methods is invalid. Compared with traditional open-set recognition tasks, Open-world WTAL (OWTAL) is challenging since not only are the annotations of unknown samples unavailable, but also the fine-grained annotations of known action instances can only be inferred ambiguously from the video category labels. To address this problem, we propose a Cascade Evidential Learning framework at an evidence level, which targets at OWTAL for the first time. Our method jointly leverages multi-scale temporal contexts and knowledge-guided prototype information to progressively collect cascade and enhanced evidence for known action, unknown action, and background separation. Extensive experiments conducted on THUMOS-14 and ActivityNet-v1.3 verify the effectiveness of our method. Besides the classification metrics adopted by previous open-set recognition methods, we also evaluate our method on localization metrics which are more reasonable for OWTAL.
****
Efficient Movie Scene Detection Using State-Space Transformers
Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey, Tony Braskich, Gedas Bertasius; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18749-18758
The ability to distinguish between different movie scenes is critical for understanding the storyline of a movie. However, accurately detecting movie scenes is often challenging as it requires the ability to reason over very long movie segments. This is in contrast to most existing video recognition models, which are typically designed for short-range video analysis. This work proposes a State-Space Transformer model that can efficiently capture dependencies in long movie videos for accurate movie scene detection. Our model, dubbed TranS4mer, is built using a novel S4A building block, which combines the strengths of structured state-space sequence (S4) and self-attention (A) layers. Given a sequence of frames divided into movie shots (uninterrupted periods where the camera position does not change), the S4A block first applies self-attention to capture short-range intra-shot dependencies. Afterward, the state-space operation in the S4A block is u

sed to aggregate long-range inter-shot cues. The final TranS4mer model, which ca
n be trained end-to-end, is obtained by stacking the S4A blocks one after the ot
her multiple times. Our proposed TranS4mer outperforms all prior methods in thre
e movie scene detection datasets, including MovieNet, BBC, and OVSD, while also
being 2x faster and requiring 3x less GPU memory than standard Transformer model
s. We will release our code and models.
****
Multispectral Video Semantic Segmentation: A Benchmark Dataset and Baseline
Wei Ji, Jingjing Li, Cheng Bian, Zongwei Zhou, Jiaying Zhao, Alan L. Yuille, Li
Cheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec
ognition (CVPR), 2023, pp. 1094-1104
Robust and reliable semantic segmentation in complex scenes is crucial for many
real-life applications such as autonomous safe driving and nighttime rescue. In
most approaches, it is typical to make use of RGB images as input. They however
work well only in preferred weather conditions; when facing adverse conditions s
uch as rainy, overexposure, or low-light, they often fail to deliver satisfactor
y results. This has led to the recent investigation into multispectral semantic
segmentation, where RGB and thermal infrared (RGBT) images are both utilized as
input. This gives rise to significantly more robust segmentation of image object
s in complex scenes and under adverse conditions. Nevertheless, the present focu
s in single RGBT image input restricts existing methods from well addressing dyn
amic real-world scenes. Motivated by the above observations, in this paper, we s
et out to address a relatively new task of semantic segmentation of multispectra
l video input, which we refer to as Multispectral Video Semantic Segmentation, o
r MVSS in short. An in-house MVSeg dataset is thus curated, consisting of 738 ca
librated RGB and thermal videos, accompanied by 3,545 fine-grained pixel-level s
emantic annotations of 26 categories. Our dataset contains a wide range of chall
enging urban scenes in both daytime and nighttime. Moreover, we propose an effec
tive MVSS baseline, dubbed MVNet, which is to our knowledge the first model to j
ointly learn semantic representations from multispectral and temporal contexts.
Comprehensive experiments are conducted using various semantic segmentation mode
ls on the MVSeg dataset. Empirically, the engagement of multispectral video inpu
t is shown to lead to significant improvement in semantic segmentation; the effe
ctiveness of our MVNet baseline has also been verified.
****
Reducing the Label Bias for Timestamp Supervised Temporal Action Segmentation
Kaiyuan Liu, Yunheng Li, Shenglan Liu, Chenwei Tan, Zihang Shao; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
 pp. 6503-6513
Timestamp supervised temporal action segmentation (TSTAS) is more cost-effective
 than fully supervised counterparts. However, previous approaches suffer from se
vere label bias due to over-reliance on sparse timestamp annotations, resulting
in unsatisfactory performance. In this paper, we propose the Debiasing-TSTAS (D-
TSTAS) framework by exploiting unannotated frames to alleviate this bias from tw
o phases: 1) Initialization. To reduce the dependencies on annotated frames, we
propose masked timestamp predictions (MTP) to ensure that initialized model capt
ures more contextual information. 2) Refinement. To overcome the limitation of t
he expressiveness from sparsely annotated timestamps, we propose a center-orient
ed timestamp expansion (CTE) approach to progressively expand pseudo-timestamp g
roups which contain semantic-rich motion representation of action segments. Then
, these pseudo-timestamp groups and the model output are used to iteratively gen
erate pseudo-labels for refining the model in a fully supervised setup. We furth
er introduce segmental confidence loss to enable the model to have high confiden
ce predictions within the pseudo-timestamp groups and more accurate action bound
aries. Our D-TSTAS outperforms the state-of-the-art TSTAS method as well as achi
eves competitive results compared with fully supervised approaches on three benc
hmark datasets.
****
Efficient Semantic Segmentation by Altering Resolutions for Compressed Videos
Yubin Hu, Yuze He, Yanghao Li, Jisheng Li, Yuxing Han, Jiangtao Wen, Yong-Jin Li

u; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22627-22637

Video semantic segmentation (VSS) is a computationally expensive task due to the per-frame prediction for videos of high frame rates. In recent work, compact models or adaptive network strategies have been proposed for efficient VSS. However, they did not consider a crucial factor that affects the computational cost from the input side: the input resolution. In this paper, we propose an altering resolution framework called AR-Seg for compressed videos to achieve efficient VSS. AR-Seg aims to reduce the computational cost by using low resolution for non-keyframes. To prevent the performance degradation caused by downsampling, we design a Cross Resolution Feature Fusion (CReFF) module, and supervise it with a novel Feature Similarity Training (FST) strategy. Specifically, CReFF first makes use of motion vectors stored in a compressed video to warp features from high-resolution keyframes to low-resolution non-keyframes for better spatial alignment, and then selectively aggregates the warped features with local attention mechanism. Furthermore, the proposed FST supervises the aggregated features with high-resolution features through an explicit similarity loss and an implicit constraint from the shared decoding layer. Extensive experiments on CamVid and Cityscapes show that AR-Seg achieves state-of-the-art performance and is compatible with different segmentation backbones. On CamVid, AR-Seg saves 67% computational cost (measured in GFLOPs) with the PSPNet18 backbone while maintaining high segmentation accuracy. Code: https://github.com/THU-LYJ-Lab/AR-Seg.

****

STAR Loss: Reducing Semantic Ambiguity in Facial Landmark Detection

Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15475-15484

Recently, deep learning-based facial landmark detection has achieved significant improvement. However, the semantic ambiguity problem degrades detection performance. Specifically, the semantic ambiguity causes inconsistent annotation and negatively affects the model's convergence, leading to worse accuracy and instability prediction. To solve this problem, we propose a Self-adapTive Ambiguity Reduction (STAR) loss by exploiting the properties of semantic ambiguity. We find that semantic ambiguity results in the anisotropic predicted distribution, which inspires us to use predicted distribution to represent semantic ambiguity. Based on this, we design the STAR loss that measures the anisotropism of the predicted distribution. Compared with the standard regression loss, STAR loss is encouraged to be small when the predicted distribution is anisotropic and thus adaptively mitigates the impact of semantic ambiguity. Moreover, we propose two kinds of eigenvalue restriction methods that could avoid both distribution's abnormal change and the model's premature convergence. Finally, the comprehensive experiments demonstrate that STAR loss outperforms the state-of-the-art methods on three benchmarks, i.e., COFW, 300W, and WFLW, with negligible computation overhead. Code is at https://github.com/ZhenglinZhou/STAR

****

A Meta-Learning Approach to Predicting Performance and Data Requirements

Achin Jain, Gurumurthy Swaminathan, Paolo Favaro, Hao Yang, Avinash Ravichandran, Hrayr Harutyunyan, Alessandro Achille, Onkar Dabeer, Bernt Schiele, Ashwin Swaminathan, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3623-3632

We propose an approach to estimate the number of samples required for a model to reach a target performance. We find that the power law, the de facto principle to estimate model performance, leads to large error when using a small dataset (e.g., 5 samples per class) for extrapolation. This is because the log-performance error against the log-dataset size follows a nonlinear progression in the few-shot regime followed by a linear progression in the high-shot regime. We introduce a novel piecewise power law (PPL) that handles the two data regimes differently. To estimate the parameters of the PPL, we introduce a random forest regressor trained via meta learning that generalizes across classification/detection tasks, ResNet/ViT based architectures, and random/pre-trained initializations. The

PPL improves the performance estimation on average by 37% across 16 classificati
on datasets and 33% across 10 detection datasets, compared to the power law. We
further extend the PPL to provide a confidence bound and use it to limit the pre
diction horizon that reduces over-estimation of data by 76% on classification an
d 91% on detection datasets.
****
Seeing What You Said: Talking Face Generation Guided by a Lip Reading Expert
Jiadong Wang, Xinyuan Qian, Malu Zhang, Robby T. Tan, Haizhou Li; Proceedings of
 the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023
, pp. 14653-14662
Talking face generation, also known as speech-to-lip generation, reconstructs fa
cial motions concerning lips given coherent speech input. The previous studies r
evealed the importance of lip-speech synchronization and visual quality. Despite
 much progress, they hardly focus on the content of lip movements i.e., the visu
al intelligibility of the spoken words, which is an important aspect of generati
on quality. To address the problem, we propose using a lip-reading expert to imp
rove the intelligibility of the generated lip regions by penalizing the incorrec
t generation results. Moreover, to compensate for data scarcity, we train the li
p-reading expert in an audio-visual self-supervised manner. With a lip-reading e
xpert, we propose a novel contrastive learning to enhance lip-speech synchroniza
tion, and a transformer to encode audio synchronically with video, while conside
ring global temporal dependency of audio. For evaluation, we propose a new strat
egy with two different lip-reading experts to measure intelligibility of the gen
erated videos. Rigorous experiments show that our proposal is superior to other
State-of-the-art (SOTA) methods, such as Wav2Lip, in reading intelligibility i.e
., over 38% Word Error Rate (WER) on LRS2 dataset and 27.8% accuracy on LRW data
set. We also achieve the SOTA performance in lip-speech synchronization and comp
arable performances in visual quality.
****
Deep Curvilinear Editing: Commutative and Nonlinear Image Manipulation for Pretr
ained Deep Generative Model
Takehiro Aoshima, Takashi Matsubara; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2023, pp. 5957-5967
Semantic editing of images is the fundamental goal of computer vision. Although
deep learning methods, such as generative adversarial networks (GANs), are capab
le of producing high-quality images, they often do not have an inherent way of e
diting generated images semantically. Recent studies have investigated a way of
manipulating the latent variable to determine the images to be generated. Howeve
r, methods that assume linear semantic arithmetic have certain limitations in te
rms of the quality of image editing, whereas methods that discover nonlinear sem
antic pathways provide non-commutative editing, which is inconsistent when appli
ed in different orders. This study proposes a novel method called deep curviline
ar editing (DeCurvEd) to determine semantic commuting vector fields on the laten
t space. We theoretically demonstrate that owing to commutativity, the editing o
f multiple attributes depends only on the quantities and not on the order. Furth
ermore, we experimentally demonstrate that compared to previous methods, the non
linear and commutative nature of DeCurvEd provides higher-quality editing.
****
Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement
Yuhui Wu, Chen Pan, Guoqing Wang, Yang Yang, Jiwei Wei, Chongyi Li, Heng Tao She
n; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogni
tion (CVPR), 2023, pp. 1662-1671
Low-light image enhancement (LLIE) investigates how to improve illumination and
produce normal-light images. The majority of existing methods improve low-light
images via a global and uniform manner, without taking into account the semantic
 information of different regions. Without semantic priors, a network may easily
 deviate from a region's original color. To address this issue, we propose a nov
el semantic-aware knowledge-guided framework (SKF) that can assist a low-light e
nhancement model in learning rich and diverse priors encapsulated in a semantic
segmentation model. We concentrate on incorporating semantic knowledge from thre

e key aspects: a semantic-aware embedding module that wisely integrates semantic priors in feature representation space, a semantic-guided color histogram loss that preserves color consistency of various instances, and a semantic-guided adversarial loss that produces more natural textures by semantic priors. Our SKF is appealing in acting as a general framework in LLIE task. Extensive experiments show that models equipped with the SKF significantly outperform the baselines on multiple datasets and our SKF generalizes to different models and scenes well. The code is available at Semantic-Aware-Low-Light-Image-Enhancement.
****
SimpSON: Simplifying Photo Cleanup With Single-Click Distracting Object Segmentation Network
Chuong Huynh, Yuqian Zhou, Zhe Lin, Connelly Barnes, Eli Shechtman, Sohrab Amirghodsi, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14518-14527
In photo editing, it is common practice to remove visual distractions to improve the overall image quality and highlight the primary subject. However, manually selecting and removing these small and dense distracting regions can be a laborious and time-consuming task. In this paper, we propose an interactive distractor selection method that is optimized to achieve the task with just a single click. Our method surpasses the precision and recall achieved by the traditional method of running panoptic segmentation and then selecting the segments containing the clicks. We also showcase how a transformer-based module can be used to identify more distracting regions similar to the user's click position. Our experiments demonstrate that the model can effectively and accurately segment unknown distracting objects interactively and in groups. By significantly simplifying the photo cleaning and retouching process, our proposed model provides inspiration for exploring rare object segmentation and group selection with a single click.
****
Learning Neural Duplex Radiance Fields for Real-Time View Synthesis
Ziyu Wan, Christian Richardt, Aljaž Boži■, Chao Li, Vijay Rengarajan, Seonghyeon Nam, Xiaoyu Xiang, Tuotuo Li, Bo Zhu, Rakesh Ranjan, Jing Liao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8307-8316
Neural radiance fields (NeRFs) enable novel view synthesis with unprecedented visual quality. However, to render photorealistic images, NeRFs require hundreds of deep multilayer perceptron (MLP) evaluations -- for each pixel. This is prohibitively expensive and makes real-time rendering infeasible, even on powerful modern GPUs. In this paper, we propose a novel approach to distill and bake NeRFs into highly efficient mesh-based neural representations that are fully compatible with the massively parallel graphics rendering pipeline. We represent scenes as neural radiance features encoded on a two-layer duplex mesh, which effectively overcomes the inherent inaccuracies in 3D surface reconstruction by learning the aggregated radiance information from a reliable interval of ray-surface intersections. To exploit local geometric relationships of nearby pixels, we leverage screen-space convolutions instead of the MLPs used in NeRFs to achieve high-quality appearance. Finally, the performance of the whole framework is further boosted by a novel multi-view distillation optimization strategy. We demonstrate the effectiveness and superiority of our approach via extensive experiments on a range of standard datasets.
****
Deep Arbitrary-Scale Image Super-Resolution via Scale-Equivariance Pursuit
Xiaohang Wang, Xuanhong Chen, Bingbing Ni, Hang Wang, Zhengyan Tong, Yutian Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1786-1795
The ability of scale-equivariance processing blocks plays a central role in arbitrary-scale image super-resolution tasks. Inspired by this crucial observation, this work proposes two novel scale-equivariant modules within a transformer-style framework to enhance arbitrary-scale image super-resolution (ASISR) performance, especially in high upsampling rate image extrapolation. In the feature extraction phase, we design a plug-in module called Adaptive Feature Extractor, which

injects explicit scale information in frequency-expanded encoding, thus achieving scale-adaption in representation learning. In the upsampling phase, a learnable Neural Kriging upsampling operator is introduced, which simultaneously encodes both relative distance (i.e., scale-aware) information as well as feature similarity (i.e., with priori learned from training data) in a bilateral manner, providing scale-encoded spatial feature fusion. The above operators are easily plugged into multiple stages of a SR network, and a recent emerging pre-training strategy is also adopted to impulse the model's performance further. Extensive experimental results have demonstrated the outstanding scale-equivariance capability offered by the proposed operators and our learning framework, with much better results than previous SOTAs at arbitrary scales for SR. Our code is available at https://github.com/neuralchen/EQSR.
****
Towards Modality-Agnostic Person Re-Identification With Descriptive Query
Cuiqun Chen, Mang Ye, Ding Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15128-15137
Person re-identification (ReID) with descriptive query (text or sketch) provides an important supplement for general image-image paradigms, which is usually studied in a single cross-modality matching manner, e.g., text-to-image or sketch-to-photo. However, without a camera-captured photo query, it is uncertain whether the text or sketch is available or not in practical scenarios. This motivates us to study a new and challenging modality-agnostic person re-identification problem. Towards this goal, we propose a unified person re-identification (UNIReID) architecture that can effectively adapt to cross-modality and multi-modality tasks. Specifically, UNIReID incorporates a simple dual-encoder with task-specific modality learning to mine and fuse visual and textual modality information. To deal with the imbalanced training problem of different tasks in UNIReID, we propose a task-aware dynamic training strategy in terms of task difficulty, adaptively adjusting the training focus. Besides, we construct three multi-modal ReID datasets by collecting the corresponding sketches from photos to support this challenging task. The experimental results on three multi-modal ReID datasets show that our UNIReID greatly improves the retrieval accuracy and generalization ability on different tasks and unseen scenarios.
****
Discriminating Known From Unknown Objects via Structure-Enhanced Recurrent Variational AutoEncoder
Aming Wu, Cheng Deng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23956-23965
Discriminating known from unknown objects is an important essential ability for human beings. To simulate this ability, a task of unsupervised out-of-distribution object detection (OOD-OD) is proposed to detect the objects that are never-seen-before during model training, which is beneficial for promoting the safe deployment of object detectors. Due to lacking unknown data for supervision, for this task, the main challenge lies in how to leverage the known in-distribution (ID) data to improve the detector's discrimination ability. In this paper, we first propose a method of Structure-Enhanced Recurrent Variational AutoEncoder (SR-VAE), which mainly consists of two dedicated recurrent VAE branches. Specifically, to boost the performance of object localization, we explore utilizing the classical Laplacian of Gaussian (LoG) operator to enhance the structure information in the extracted low-level features. Meanwhile, we design a VAE branch that recurrently generates the augmentation of the classification features to strengthen the discrimination ability of the object classifier. Finally, to alleviate the impact of lacking unknown data, another cycle-consistent conditional VAE branch is proposed to synthesize virtual OOD features that deviate from the distribution of ID features, which improves the capability of distinguishing OOD objects. In the experiments, our method is evaluated on OOD-OD, open-vocabulary detection, and incremental object detection. The significant performance gains over baselines show the superiorities of our method. The code will be released at https://github.com/AmingWu/SR-VAE.
****

Occlusion-Free Scene Recovery via Neural Radiance Fields
Chengxuan Zhu, Renjie Wan, Yunkai Tang, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20722-20731
Our everyday lives are filled with occlusions that we strive to see through. By aggregating desired background information from different viewpoints, we can easily eliminate such occlusions without any external occlusion-free supervision. Though several occlusion removal methods have been proposed to empower machine vision systems with such ability, their performances are still unsatisfactory due to reliance on external supervision. We propose a novel method for occlusion removal by directly building a mapping between position and viewing angles and the corresponding occlusion-free scene details leveraging Neural Radiance Fields (NeRF). We also develop an effective scheme to jointly optimize camera parameters and scene reconstruction when occlusions are present. An additional depth constraint is applied to supervise the entire optimization without labeled external data for training. The experimental results on existing and newly collected datasets validate the effectiveness of our method.
****
OmniAL: A Unified CNN Framework for Unsupervised Anomaly Localization
Ying Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3924-3933
Unsupervised anomaly localization and detection is crucial for industrial manufacturing processes due to the lack of anomalous samples. Recent unsupervised advances on industrial anomaly detection achieve high performance by training separate models for many different categories. The model storage and training time cost of this paradigm is high. Moreover, the setting of one-model-N-classes leads to fearful degradation of existing methods. In this paper, we propose a unified CNN framework for unsupervised anomaly localization, named OmniAL. This method conquers aforementioned problems by improving anomaly synthesis, reconstruction and localization. To prevent the model learning identical reconstruction, it trains the model with proposed panel-guided synthetic anomaly data rather than directly using normal data. It increases anomaly reconstruction error for multi-class distribution by using a network that is equipped with proposed Dilated Channel and Spatial Attention (DCSA) blocks. To better localize the anomaly regions, it employs proposed DiffNeck between reconstruction and localization sub-networks to explore multi-level differences. Experiments on 15-class MVTecAD and 12-class VisA datasets verify the advantage of proposed OmniAL that surpasses the state-of-the-art of unified models. On 15-class-MVTecAD/12-class-VisA, its single unified model achieves 97.2/87.8 image-AUROC, 98.3/96.6 pixel-AUROC and 73.4/41.7 pixel-AP for anomaly detection and localization respectively. Besides that, we make the first attempt to conduct a comprehensive study on the robustness of unsupervised anomaly localization and detection methods against different level adversarial attacks. Experiential results show OmniAL has good application prospects for its superior performance.
****
An In-Depth Exploration of Person Re-Identification and Gait Recognition in Cloth-Changing Conditions
Weijia Li, Saihui Hou, Chunjie Zhang, Chunshui Cao, Xu Liu, Yongzhen Huang, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13824-13833
The target of person re-identification (ReID) and gait recognition is consistent, that is to match the target pedestrian under surveillance cameras. For the cloth-changing problem, video-based ReID is rarely studied due to the lack of a suitable cloth-changing benchmark, and gait recognition is often researched under controlled conditions. To tackle this problem, we propose a Cloth-Changing benchmark for Person re-identification and Gait recognition (CCPG). It is a cloth-changing dataset, and there are several highlights in CCPG, (1) it provides 200 identities and over 16K sequences are captured indoors and outdoors, (2) each identity has seven different cloth-changing statuses, which is hardly seen in previous datasets, (3) RGB and silhouettes version data are both available for research

purposes. Moreover, aiming to investigate the cloth-changing problem systematically, comprehensive experiments are conducted on video-based ReID and gait recognition methods. The experimental results demonstrate the superiority of ReID and gait recognition separately in different cloth-changing conditions and suggest that gait recognition is a potential solution for addressing the cloth-changing problem. Our dataset will be available at https://github.com/BNU-IVC/CCPG.
****
Visual Exemplar Driven Task-Prompting for Unified Perception in Autonomous Driving
Xiwen Liang, Minzhe Niu, Jianhua Han, Hang Xu, Chunjing Xu, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9611-9621
Multi-task learning has emerged as a powerful paradigm to solve a range of tasks simultaneously with good efficiency in both computation resources and inference time. However, these algorithms are designed for different tasks mostly not within the scope of autonomous driving, thus making it hard to compare multi-task methods in autonomous driving. Aiming to enable the comprehensive evaluation of present multi-task learning methods in autonomous driving, we extensively investigate the performance of popular multi-task methods on the large-scale driving dataset, which covers four common perception tasks, i.e., object detection, semantic segmentation, drivable area segmentation, and lane detection. We provide an in-depth analysis of current multi-task learning methods under different common settings and find out that the existing methods make progress but there is still a large performance gap compared with single-task baselines. To alleviate this dilemma in autonomous driving, we present an effective multi-task framework, VE-Prompt, which introduces visual exemplars via task-specific prompting to guide the model toward learning high-quality task-specific representations. Specifically, we generate visual exemplars based on bounding boxes and color-based markers, which provide accurate visual appearances of target categories and further mitigate the performance gap. Furthermore, we bridge transformer-based encoders and convolutional layers for efficient and accurate unified perception in autonomous driving. Comprehensive experimental results on the diverse self-driving dataset BDD100K show that the VE-Prompt improves the multi-task baseline and further surpasses single-task models.
****
Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation
Mayu Otani, Riku Togashi, Yu Sawai, Ryosuke Ishigami, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, Shin’ichi Satoh; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14277-14286
Human evaluation is critical for validating the performance of text-to-image generative models, as this highly cognitive process requires deep comprehension of text and images. However, our survey of 37 recent papers reveals that many works rely solely on automatic measures (e.g., FID) or perform poorly described human evaluations that are not reliable or repeatable. This paper proposes a standardized and well-defined human evaluation protocol to facilitate verifiable and reproducible human evaluation in future works. In our pilot data collection, we experimentally show that the current automatic measures are incompatible with human perception in evaluating the performance of the text-to-image generation results. Furthermore, we provide insights for designing human evaluation experiments reliably and conclusively. Finally, we make several resources publicly available to the community to facilitate easy and fast implementations.
****
Semi-Supervised Domain Adaptation With Source Label Adaptation
Yu-Chu Yu, Hsuan-Tien Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24100-24109
Semi-Supervised Domain Adaptation (SSDA) involves learning to classify unseen target data with a few labeled and lots of unlabeled target data, along with many labeled source data from a related domain. Current SSDA approaches usually aim at aligning the target data to the labeled source data with feature space mapping and pseudo-label assignments. Nevertheless, such a source-oriented model can so

metimes align the target data to source data of the wrong classes, degrading the classification performance. This paper presents a novel source-adaptive paradigm that adapts the source data to match the target data. Our key idea is to view the source data as a noisily-labeled version of the ideal target data. Then, we propose an SSDA model that cleans up the label noise dynamically with the help of a robust cleaner component designed from the target perspective. Since the paradigm is very different from the core ideas behind existing SSDA approaches, our proposed model can be easily coupled with them to improve their performance. Empirical results on two state-of-the-art SSDA approaches demonstrate that the proposed model effectively cleans up the noise within the source labels and exhibits superior performance over those approaches across benchmark datasets. Our code is available at https://github.com/chu0802/SLA.
****
Range-Nullspace Video Frame Interpolation With Focalized Motion Estimation
Zhiyang Yu, Yu Zhang, Dongqing Zou, Xijun Chen, Jimmy S. Ren, Shunqing Ren; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22159-22168
Continuous-time video frame interpolation is a fundamental technique in computer vision for its flexibility in synthesizing motion trajectories and novel video frames at arbitrary intermediate time steps. Yet, how to infer accurate intermediate motion and synthesize high-quality video frames are two critical challenges. In this paper, we present a novel VFI framework with improved treatment for these challenges. To address the former, we propose focalized trajectory fitting, which performs confidence-aware motion trajectory estimation by learning to pay focus to reliable optical flow candidates while suppressing the outliers. The second is range-nullspace synthesis, a novel frame renderer cast as solving an ill-posed problem addressed by learning decoupled components in orthogonal subspaces. The proposed framework sets new records on 7 of 10 public VFI benchmarks.
****
FlowGrad: Controlling the Output of Generative ODEs With Gradients
Xingchao Liu, Lemeng Wu, Shujian Zhang, Chengyue Gong, Wei Ping, Qiang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24335-24344
Generative modeling with ordinary differential equations (ODEs) has achieved fantastic results on a variety of applications. Yet, few works have focused on controlling the generated content of a pre-trained ODE-based generative model. In this paper, we propose to optimize the output of ODE models according to a guidance function to achieve controllable generation. We point out that, the gradients can be efficiently back-propagated from the output to any intermediate time steps on the ODE trajectory, by decomposing the back-propagation and computing vector-Jacobian products. To further accelerate the computation of the back-propagation, we propose to use a non-uniform discretization to approximate the ODE trajectory, where we measure how straight the trajectory is and gather the straight parts into one discretization step. This allows us to save 90% of the back-propagation time with ignorable error. Our framework, named FlowGrad, outperforms the state-of-the-art baselines on text-guided image manipulation. Moreover, FlowGrad enables us to find global semantic directions in frozen ODE-based generative models that can be used to manipulate new images without extra optimization.
****
Learning Weather-General and Weather-Specific Features for Image Restoration Under Multiple Adverse Weather Conditions
Yurui Zhu, Tianyu Wang, Xueyang Fu, Xuanyu Yang, Xin Guo, Jifeng Dai, Yu Qiao, Xiaowei Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21747-21758
Image restoration under multiple adverse weather conditions aims to remove weather-related artifacts by using the single set of network parameters. In this paper, we find that distorted images under different weather conditions contain general characteristics as well as their specific characteristics. Inspired by this observation, we design an efficient unified framework with a two-stage training strategy to explore the weather-general and weather-specific features. The first

training stage aims to learn the weather-general features by taking the images under various weather conditions as the inputs and outputting the coarsely resto red results. The second training stage aims to learn to adaptively expand the sp ecific parameters for each weather type in the deep model, where requisite posit ions for expansion of weather-specific parameters are learned automatically. Hen ce, we can obtain an efficient and unified model for image restoration under mul tiple adverse weather conditions. Moreover, we build the first real-world benchm ark dataset with multiple weather conditions to better deal with real-world weat her scenarios. Experimental results show that our method achieves superior perfo rmance on all the synthetic and real-world benchmark datasets.
****
Generalized Deep 3D Shape Prior via Part-Discretized Diffusion Process
Yuhan Li, Yishun Dou, Xuanhong Chen, Bingbing Ni, Yilin Sun, Yutian Liu, Fuzhen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2023, pp. 16784-16794
We develop a generalized 3D shape generation prior model, tailored for multiple 3D tasks including unconditional shape generation, point cloud completion, and c ross-modality shape generation, etc. On one hand, to precisely capture local fin e detailed shape information, a vector quantized variational autoencoder (VQ-VAE) is utilized to index local geometry from a compactly learned codebook based on a broad set of task training data. On the other hand, a discrete diffusion gene rator is introduced to model the inherent structural dependencies among differen t tokens. In the meantime, a multi-frequency fusion module (MFM) is developed to suppress high-frequency shape feature fluctuations, guided by multi-frequency c ontextual information. The above designs jointly equip our proposed 3D shape pri or model with high-fidelity, diverse features as well as the capability of cross -modality alignment, and extensive experiments have demonstrated superior perfor mances on various 3D shape generation tasks.
****
Conflict-Based Cross-View Consistency for Semi-Supervised Semantic Segmentation
Zicheng Wang, Zhen Zhao, Xiaoxia Xing, Dong Xu, Xiangyu Kong, Luping Zhou; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2023, pp. 19585-19595
Semi-supervised semantic segmentation (SSS) has recently gained increasing resea rch interest as it can reduce the requirement for large-scale fully-annotated tr aining data. The current methods often suffer from the confirmation bias from th e pseudo-labelling process, which can be alleviated by the co-training framework . The current co-training-based SSS methods rely on hand-crafted perturbations t o prevent the different sub-nets from collapsing into each other, but these arti ficial perturbations cannot lead to the optimal solution. In this work, we propo se a new conflict-based cross-view consistency (CCVC) method based on a two-bran ch co-training framework which aims at enforcing the two sub-nets to learn infor mative features from irrelevant views. In particular, we first propose a new cro ss-view consistency (CVC) strategy that encourages the two sub-nets to learn dis tinct features from the same input by introducing a feature discrepancy loss, wh ile these distinct features are expected to generate consistent prediction score s of the input. The CVC strategy helps to prevent the two sub-nets from stepping into the collapse. In addition, we further propose a conflict-based pseudo-labe lling (CPL) method to guarantee the model will learn more useful information fro m conflicting predictions, which will lead to a stable training process. We vali date our new CCVC approach on the SSS benchmark datasets where our method achiev es new state-of-the-art performance. Our code is available at https://github.com /xiaoyao3302/CCVC.
****
Learning a 3D Morphable Face Reflectance Model From Low-Cost Data
Yuxuan Han, Zhibo Wang, Feng Xu; Proceedings of the IEEE/CVF Conference on Compu ter Vision and Pattern Recognition (CVPR), 2023, pp. 8598-8608
Modeling non-Lambertian effects such as facial specularity leads to a more reali stic 3D Morphable Face Model. Existing works build parametric models for diffuse and specular albedo using Light Stage data. However, only diffuse and specular

albedo cannot determine the full BRDF. In addition, the requirement of Light Stage data is hard to fulfill for the research communities. This paper proposes the first 3D morphable face reflectance model with spatially varying BRDF using only low-cost publicly-available data. We apply linear shiness weighting into parametric modeling to represent spatially varying specular intensity and shiness. Then an inverse rendering algorithm is developed to reconstruct the reflectance parameters from non-Light Stage data, which are used to train an initial morphable reflectance model. To enhance the model's generalization capability and expressive power, we further propose an update-by-reconstruction strategy to finetune it on an in-the-wild dataset. Experimental results show that our method obtains decent rendering results with plausible facial specularities. Our code is released at https://yxuhan.github.io/ReflectanceMM/index.html.
****
SCoDA: Domain Adaptive Shape Completion for Real Scans
Yushuang Wu, Zizheng Yan, Ce Chen, Lai Wei, Xiao Li, Guanbin Li, Yihao Li, Shuguang Cui, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17630-17641
3D shape completion from point clouds is a challenging task, especially from scans of real-world objects. Considering the paucity of 3D shape ground truths for real scans, existing works mainly focus on benchmarking this task on synthetic data, e.g. 3D computer-aided design models. However, the domain gap between synthetic and real data limits the generalizability of these methods. Thus, we propose a new task, SCoDA, for the domain adaptation of real scan shape completion from synthetic data. A new dataset, ScanSalon, is contributed with a bunch of elaborate 3D models created by skillful artists according to scans. To address this new task, we propose a novel cross-domain feature fusion method for knowledge transfer and a novel volume-consistent self-training framework for robust learning from real data. Extensive experiments prove our method is effective to bring an improvement of 6% 7% mIoU.
****
Recurrent Homography Estimation Using Homography-Guided Image Warping and Focus Transformer
Si-Yuan Cao, Runmin Zhang, Lun Luo, Beinan Yu, Zehua Sheng, Junwei Li, Hui-Liang Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9833-9842
We propose the Recurrent homography estimation framework using Homography-guided image Warping and Focus transformer (FocusFormer), named RHWF. Both being appropriately absorbed into the recurrent framework, the homography-guided image warping progressively enhances the feature consistency and the attention-focusing mechanism in FocusFormer aggregates the intra-inter correspondence in a global->nonlocal->local manner. Thanks to the above strategies, RHWF ranks top in accuracy on a variety of datasets, including the challenging cross-resolution and cross-modal ones. Meanwhile, benefiting from the recurrent framework, RHWF achieves parameter efficiency despite the transformer architecture. Compared to previous state-of-the-art approaches LocalTrans and IHN, RHWF reduces the mean average corner error (MACE) by about 70% and 38.1% on the MSCOCO dataset, while saving the parameter costs by 86.5% and 24.6%. Similar to the previous works, RHWF can also be arranged in 1-scale for efficiency and 2-scale for accuracy, with the 1-scale RHWF already outperforming most of the previous methods. Source code is available at https://github.com/imdumpl78/RHWF.
****
I2-SDF: Intrinsic Indoor Scene Reconstruction and Editing via Raytracing in Neural SDFs
Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, Rui Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12489-12498
In this work, we present I^2-SDF, a new method for intrinsic indoor scene reconstruction and editing using differentiable Monte Carlo raytracing on neural signed distance fields (SDFs). Our holistic neural SDF-based framework jointly recovers the underlying shapes, incident radiance and materials from multi-view images

. We introduce a novel bubble loss for fine-grained small objects and error-guided adaptive sampling scheme to largely improve the reconstruction quality on large-scale indoor scenes. Further, we propose to decompose the neural radiance field into spatially-varying material of the scene as a neural field through surface-based, differentiable Monte Carlo raytracing and emitter semantic segmentations, which enables physically based and photorealistic scene relighting and editing applications. Through a number of qualitative and quantitative experiments, we demonstrate the superior quality of our method on indoor scene reconstruction, novel view synthesis, and scene editing compared to state-of-the-art baselines. Our project page is at https://jingsenzhu.github.io/i2-sdf.
****
DLBD: A Self-Supervised Direct-Learned Binary Descriptor
Bin Xiao, Yang Hu, Bo Liu, Xiuli Bi, Weisheng Li, Xinbo Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15846-15855
For learning-based binary descriptors, the binarization process has not been well addressed. The reason is that the binarization blocks gradient back-propagation. Existing learning-based binary descriptors learn real-valued output, and then it is converted to binary descriptors by their proposed binarization processes. Since their binarization processes are not a component of the network, the learning-based binary descriptor cannot fully utilize the advances of deep learning. To solve this issue, we propose a model-agnostic plugin binary transformation layer (BTL), making the network directly generate binary descriptors. Then, we present the first self-supervised, direct-learned binary descriptor, dubbed DLBD. Furthermore, we propose ultra-wide temperature-scaled cross-entropy loss to adjust the distribution of learned descriptors in a larger range. Experiments demonstrate that the proposed BTL can substitute the previous binarization process. Our proposed DLBD outperforms SOTA on different tasks such as image retrieval and classification.
****
Fuzzy Positive Learning for Semi-Supervised Semantic Segmentation
Pengchong Qiao, Zhidan Wei, Yu Wang, Zhennan Wang, Guoli Song, Fan Xu, Xiangyang Ji, Chang Liu, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15465-15474
Semi-supervised learning (SSL) essentially pursues class boundary exploration with less dependence on human annotations. Although typical attempts focus on ameliorating the inevitable error-prone pseudo-labeling, we think differently and resort to exhausting informative semantics from multiple probably correct candidate labels. In this paper, we introduce Fuzzy Positive Learning (FPL) for accurate SSL semantic segmentation in a plug-and-play fashion, targeting adaptively encouraging fuzzy positive predictions and suppressing highly-probable negatives. Being conceptually simple yet practically effective, FPL can remarkably alleviate interference from wrong pseudo labels and progressively achieve clear pixel-level semantic discrimination. Concretely, our FPL approach consists of two main components, including fuzzy positive assignment (FPA) to provide an adaptive number of labels for each pixel and fuzzy positive regularization (FPR) to restrict the predictions of fuzzy positive categories to be larger than the rest under different perturbations. Theoretical analysis and extensive experiments on Cityscapes and VOC 2012 with consistent performance gain justify the superiority of our approach. Codes are available in https://github.com/qpc1611094/FPL.
****
Canonical Fields: Self-Supervised Learning of Pose-Canonicalized Neural Fields
Rohith Agaram, Shaurya Dewan, Rahul Sajnani, Adrien Poulenard, Madhava Krishna, Srinath Sridhar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4500-4510
Coordinate-based implicit neural networks, or neural fields, have emerged as useful representations of shape and appearance in 3D computer vision. Despite advances however, it remains challenging to build neural fields for categories of objects without datasets like ShapeNet that provide "canonicalized" object instances that are consistently aligned for their 3D position and orientation (pose). We

 present Canonical Field Network (CaFi-Net), a self-supervised method to canonicalize the 3D pose of instances from an object category represented as neural fields, specifically neural radiance fields (NeRFs). CaFi-Net directly learns from continuous and noisy radiance fields using a Siamese network architecture that is designed to extract equivariant field features for category-level canonicalization. During inference, our method takes pre-trained neural radiance fields of novel object instances at arbitrary 3D pose, and estimates a canonical field with consistent 3D pose across the entire category. Extensive experiments on a new dataset of 1300 NeRF models across 13 object categories show that our method matches or exceeds the performance of 3D point cloud-based methods.
****
TransFlow: Transformer As Flow Learner
Yawen Lu, Qifan Wang, Siqi Ma, Tong Geng, Yingjie Victor Chen, Huaijin Chen, Dongfang Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18063-18073
Optical flow is an indispensable building block for various important computer vision tasks, including motion estimation, object tracking, and disparity measurement. In this work, we propose TransFlow, a pure transformer architecture for optical flow estimation. Compared to dominant CNN-based methods, TransFlow demonstrates three advantages. First, it provides more accurate correlation and trustworthy matching in flow estimation by utilizing spatial self-attention and cross-attention mechanisms between adjacent frames to effectively capture global dependencies; Second, it recovers more compromised information (e.g., occlusion and motion blur) in flow estimation through long-range temporal association in dynamic scenes; Third, it enables a concise self-learning paradigm and effectively eliminate the complex and laborious multi-stage pre-training procedures. We achieve the state-of-the-art results on the Sintel, KITTI-15, as well as several downstream tasks, including video object detection, interpolation and stabilization. For its efficacy, we hope TransFlow could serve as a flexible baseline for optical flow estimation.
****
Multi-View Inverse Rendering for Large-Scale Real-World Indoor Scenes
Zhen Li, Lingli Wang, Mofang Cheng, Cihui Pan, Jiaqi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12499-12509
We present a efficient multi-view inverse rendering method for large-scale real-world indoor scenes that reconstructs global illumination and physically-reasonable SVBRDFs. Unlike previous representations, where the global illumination of large scenes is simplified as multiple environment maps, we propose a compact representation called Texture-based Lighting (TBL). It consists of 3D mesh and HDR textures, and efficiently models direct and infinite-bounce indirect lighting of the entire large scene. Based on TBL, we further propose a hybrid lighting representation with precomputed irradiance, which significantly improves the efficiency and alleviates the rendering noise in the material optimization. To physically disentangle the ambiguity between materials, we propose a three-stage material optimization strategy based on the priors of semantic segmentation and room segmentation. Extensive experiments show that the proposed method outperforms the state-of-the-art quantitatively and qualitatively, and enables physically-reasonable mixed-reality applications such as material editing, editable novel view synthesis and relighting. The project page is at https://lzleejean.github.io/TexIR.
****
AutoFocusFormer: Image Segmentation off the Grid
Chen Ziwen, Kaushik Patnaik, Shuangfei Zhai, Alvin Wan, Zhile Ren, Alexander G. Schwing, Alex Colburn, Li Fuxin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18227-18236
Real world images often have highly imbalanced content density. Some areas are very uniform, e.g., large patches of blue sky, while other areas are scattered with many small objects. Yet, the commonly used successive grid downsampling strategy in convolutional deep networks treats all areas equally. Hence, small object

s are represented in very few spatial locations, leading to worse results in tasks such as segmentation. Intuitively, retaining more pixels representing small objects during downsampling helps to preserve important information. To achieve this, we propose AutoFocusFormer (AFF), a local-attention transformer image recognition backbone, which performs adaptive downsampling by learning to retain the most important pixels for the task. Since adaptive downsampling generates a set of pixels irregularly distributed on the image plane, we abandon the classic grid structure. Instead, we develop a novel point-based local attention block, facilitated by a balanced clustering module and a learnable neighborhood merging module, which yields representations for our point-based versions of state-of-the-art segmentation heads. Experiments show that our AutoFocusFormer (AFF) improves significantly over baseline models of similar sizes.
****
Boosting Transductive Few-Shot Fine-Tuning With Margin-Based Uncertainty Weighting and Probability Regularization
Ran Tao, Hao Chen, Marios Savvides; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15752-15761
Few-Shot Learning (FSL) has been rapidly developed in recent years, potentially eliminating the requirement for significant data acquisition. Few-shot fine-tuning has been demonstrated to be practically efficient and helpful, especially for out-of-distribution datum. In this work, we first observe that the few-shot fine-tuned methods are learned with the imbalanced class marginal distribution. This observation further motivates us to propose the Transductive Fine-tuning with Margin-based uncertainty weighting and Probability regularization (TF-MP), which learns a more balanced class marginal distribution. We first conduct sample weighting on the testing data with margin-based uncertainty scores and further regularize each test sample's categorical probability. TF-MP achieves state-of-the-art performance on in- / out-of-distribution evaluations of Meta-Dataset and surpasses previous transductive methods by a large margin.
****
SMPConv: Self-Moving Point Representations for Continuous Convolution
Sanghyeon Kim, Eunbyung Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10289-10299
Continuous convolution has recently gained prominence due to its ability to handle irregularly sampled data and model long-term dependency. Also, the promising experimental results of using large convolutional kernels have catalyzed the development of continuous convolution since they can construct large kernels very efficiently. Leveraging neural networks, more specifically multilayer perceptrons (MLPs), is by far the most prevalent approach to implementing continuous convolution. However, there are a few drawbacks, such as high computational costs, complex hyperparameter tuning, and limited descriptive power of filters. This paper suggests an alternative approach to building a continuous convolution without neural networks, resulting in more computationally efficient and improved performance. We present self-moving point representations where weight parameters freely move, and interpolation schemes are used to implement continuous functions. When applied to construct convolutional kernels, the experimental results have shown improved performance with drop-in replacement in the existing frameworks. Due to its lightweight structure, we are first to demonstrate the effectiveness of continuous convolution in a large-scale setting, e.g., ImageNet, presenting the improvements over the prior arts. Our code is available on https://github.com/sangnekim/SMPConv
****
CLIP2Protect: Protecting Facial Privacy Using Text-Guided Makeup via Adversarial Latent Search
Fahad Shamshad, Muzammal Naseer, Karthik Nandakumar; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20595-20605
The success of deep learning based face recognition systems has given rise to serious privacy concerns due to their ability to enable unauthorized tracking of users in the digital world. Existing methods for enhancing privacy fail to genera

te naturalistic' images that can protect facial privacy without compromising use
r experience. We propose a novel two-step approach for facial privacy protection
 that relies on finding adversarial latent codes in the low-dimensional manifold
 of a pretrained generative model. The first step inverts the given face image i
nto the latent space and finetunes the generative model to achieve an accurate r
econstruction of the given image from its latent code. This step produces a good
 initialization, aiding the generation of high-quality faces that resemble the g
iven identity. Subsequently, user defined makeup text prompts and identity-prese
rving regularization are used to guide the search for adversarial codes in the l
atent space. Extensive experiments demonstrate that faces generated by our appro
ach have stronger black-box transferability with an absolute gain of 12.06% over
 the state-of-the-art facial privacy protection approach under the face verifica
tion task. Finally, we demonstrate the effectiveness of the proposed approach fo
r commercial face recognition systems. Our code is available at https://github.c
om/fahadshamshad/Clip2Protect.
****
Improving Weakly Supervised Temporal Action Localization by Bridging Train-Test
Gap in Pseudo Labels
Jingqiu Zhou, Linjiang Huang, Liang Wang, Si Liu, Hongsheng Li; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 23003-23012
The task of weakly supervised temporal action localization targets at generating
 temporal boundaries for actions of interest, meanwhile the action category shou
ld also be classified. Pseudo-label-based methods, which serve as an effective s
olution, have been widely studied recently. However, existing methods generate p
seudo labels during training and make predictions during testing under different
 pipelines or settings, resulting in a gap between training and testing. In this
 paper, we propose to generate high-quality pseudo labels from the predicted act
ion boundaries. Nevertheless, we note that existing post-processing, like NMS, w
ould lead to information loss, which is insufficient to generate high-quality ac
tion boundaries. More importantly, transforming action boundaries into pseudo la
bels is quite challenging, since the predicted action instances are generally ov
erlapped and have different confidence scores. Besides, the generated pseudo-lab
els can be fluctuating and inaccurate at the early stage of training. It might r
epeatedly strengthen the false predictions if there is no mechanism to conduct s
elf-correction. To tackle these issues, we come up with an effective pipeline fo
r learning better pseudo labels. Firstly, we propose a Gaussian weighted fusion
module to preserve information of action instances and obtain high-quality actio
n boundaries. Second, we formulate the pseudo-label generation as an optimizatio
n problem under the constraints in terms of the confidence scores of action inst
ances. Finally, we introduce the idea of Delta pseudo labels, which enables the
model with the ability of self-correction. Our method achieves superior performa
nce to existing methods on two benchmarks, THUMOS14 and ActivityNet1.3, achievin
g gains of 1.9% on THUMOS14 and 3.7% on ActivityNet1.3 in terms of average mAP.
****
PRISE: Demystifying Deep Lucas-Kanade With Strongly Star-Convex Constraints for
Multimodel Image Alignment
Yiqing Zhang, Xinming Huang, Ziming Zhang; Proceedings of the IEEE/CVF Conferenc
e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13187-13197
The Lucas-Kanade (LK) method is a classic iterative homography estimation algori
thm for image alignment, but often suffers from poor local optimality especially
 when image pairs have large distortions. To address this challenge, in this pap
er we propose a novel Deep Star-Convexified Lucas-Kanade (PRISE) method for mul
timodel image alignment by introducing strongly star-convex constraints into the
 optimization problem. Our basic idea is to enforce the neural network to approx
imately learn a star-convex loss landscape around the ground truth give any data
 to facilitate the convergence of the LK method to the ground truth through the
high dimensional space defined by the network. This leads to a minimax learning
problem, with contrastive (hinge) losses due to the definition of strong star-co
nvexity that are appended to the original loss for training. We also provide an

efficient sampling based algorithm to leverage the training cost, as well as some analysis on the quality of the solutions from PRISE. We further evaluate our approach on benchmark datasets such as MSCOCO, GoogleEarth, and GoogleMap, and demonstrate state-of-the-art results, especially for small pixel errors. Demo code is attached.
****
Learning To Exploit Temporal Structure for Biomedical Vision-Language Processing
Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, Ozan Oktay; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15016-15027
Self-supervised learning in vision--language processing (VLP) exploits semantic alignment between imaging and text modalities. Prior work in biomedical VLP has mostly relied on the alignment of single image and report pairs even though clinical notes commonly refer to prior images. This does not only introduce poor alignment between the modalities but also a missed opportunity to exploit rich self-supervision through existing temporal content in the data. In this work, we explicitly account for prior images and reports when available during both training and fine-tuning. Our approach, named BioViL-T, uses a CNN--Transformer hybrid multi-image encoder trained jointly with a text model. It is designed to be versatile to arising challenges such as pose variations and missing input images across time. The resulting model excels on downstream tasks both in single- and multi-image setups, achieving state-of-the-art (SOTA) performance on (I) progression classification, (II) phrase grounding, and (III) report generation, whilst offering consistent improvements on disease classification and sentence-similarity tasks. We release a novel multi-modal temporal benchmark dataset, CXR-T, to quantify the quality of vision--language representations in terms of temporal semantics. Our experimental results show the significant advantages of incorporating prior images and reports to make most use of the data.
****
Simple Cues Lead to a Strong Multi-Object Tracker
Jenny Seidenschwarz, Guillem Brasó, Víctor Castro Serrano, Ismail Elezi, Laura Leal-Taixé; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13813-13823
For a long time, the most common paradigm in MultiObject Tracking was tracking-by-detection (TbD), where objects are first detected and then associated over video frames. For association, most models resourced to motion and appearance cues, e.g., re-identification networks. Recent approaches based on attention propose to learn the cues in a data-driven manner, showing impressive results. In this paper, we ask ourselves whether simple good old TbD methods are also capable of achieving the performance of end-to-end models. To this end, we propose two key ingredients that allow a standard re-identification network to excel at appearance-based tracking. We extensively analyse its failure cases, and show that a combination of our appearance features with a simple motion model leads to strong tracking results. Our tracker generalizes to four public datasets, namely MOT17, MOT20, BDD100k, and DanceTrack, achieving state-ofthe-art performance. https://github.com/dvl-tum/GHOST
****
Marching-Primitives: Shape Abstraction From Signed Distance Function
Weixiao Liu, Yuwei Wu, Sipu Ruan, Gregory S. Chirikjian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8771-8780
Representing complex objects with basic geometric primitives has long been a topic in computer vision. Primitive-based representations have the merits of compactness and computational efficiency in higher-level tasks such as physics simulation, collision checking, and robotic manipulation. Unlike previous works which extract polygonal meshes from a signed distance function (SDF), in this paper, we present a novel method, named Marching-Primitives, to obtain a primitive-based abstraction directly from an SDF. Our method grows geometric primitives (such as

superquadrics) iteratively by analyzing the connectivity of voxels while marching at different levels of signed distance. For each valid connected volume of interest, we march on the scope of voxels from which a primitive is able to be extracted in a probabilistic sense and simultaneously solve for the parameters of the primitive to capture the underlying local geometry. We evaluate the performance of our method on both synthetic and real-world datasets. The results show that the proposed method outperforms the state-of-the-art in terms of accuracy, and is directly generalizable among different categories and scales. The code is open-sourced at https://github.com/ChirikjianLab/Marching-Primitives.git.
****

BiasAdv: Bias-Adversarial Augmentation for Model Debiasing

Jongin Lim, Youngdong Kim, Byungjai Kim, Chanho Ahn, Jinwoo Shin, Eunho Yang, Seungju Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3832-3841

Neural networks are often prone to bias toward spurious correlations inherent in a dataset, thus failing to generalize unbiased test criteria. A key challenge to resolving the issue is the significant lack of bias-conflicting training data (i.e., samples without spurious correlations). In this paper, we propose a novel data augmentation approach termed Bias-Adversarial augmentation (BiasAdv) that supplements bias-conflicting samples with adversarial images. Our key idea is that an adversarial attack on a biased model that makes decisions based on spurious correlations may generate synthetic bias-conflicting samples, which can then be used as augmented training data for learning a debiased model. Specifically, we formulate an optimization problem for generating adversarial images that attack the predictions of an auxiliary biased model without ruining the predictions of the desired debiased model. Despite its simplicity, we find that BiasAdv can generate surprisingly useful synthetic bias-conflicting samples, allowing the debiased model to learn generalizable representations. Furthermore, BiasAdv does not require any bias annotations or prior knowledge of the bias type, which enables its broad applicability to existing debiasing methods to improve their performances. Our extensive experimental results demonstrate the superiority of BiasAdv, achieving state-of-the-art performance on four popular benchmark datasets across various bias domains.
****

CDDFuse: Correlation-Driven Dual-Branch Feature Decomposition for Multi-Modality Image Fusion

Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5906-5916

Multi-modality (MM) image fusion aims to render fused images that maintain the merits of different modalities, e.g., functional highlight and detailed textures. To tackle the challenge in modeling cross-modality features and decomposing desirable modality-specific and modality-shared features, we propose a novel Correlation-Driven feature Decomposition Fusion (CDDFuse) network. Firstly, CDDFuse uses Restormer blocks to extract cross-modality shallow features. We then introduce a dual-branch Transformer-CNN feature extractor with Lite Transformer (LT) blocks leveraging long-range attention to handle low-frequency global features and Invertible Neural Networks (INN) blocks focusing on extracting high-frequency local information. A correlation-driven loss is further proposed to make the low-frequency features correlated while the high-frequency features uncorrelated based on the embedded information. Then, the LT-based global fusion and INN-based local fusion layers output the fused image. Extensive experiments demonstrate that our CDDFuse achieves promising results in multiple fusion tasks, including infrared-visible image fusion and medical image fusion. We also show that CDDFuse can boost the performance in downstream infrared-visible semantic segmentation and object detection in a unified benchmark. The code is available at https://github.com/Zhaozixiang1228/MMIF-CDDFuse.
****

Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval

Ding Jiang, Mang Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2787-2797

Text-to-image person retrieval aims to identify the target person based on a given textual description query. The primary challenge is to learn the mapping of visual and textual modalities into a common latent space. Prior works have attempted to address this challenge by leveraging separately pre-trained unimodal models to extract visual and textual features. However, these approaches lack the necessary underlying alignment capabilities required to match multimodal data effectively. Besides, these works use prior information to explore explicit part alignments, which may lead to the distortion of intra-modality information. To alleviate these issues, we present IRRA: a cross-modal Implicit Relation Reasoning and Aligning framework that learns relations between local visual-textual tokens and enhances global image-text matching without requiring additional prior supervision. Specifically, we first design an Implicit Relation Reasoning module in a masked language modeling paradigm. This achieves cross-modal interaction by integrating the visual cues into the textual tokens with a cross-modal multimodal interaction encoder. Secondly, to globally align the visual and textual embeddings, Similarity Distribution Matching is proposed to minimize the KL divergence between image-text similarity distributions and the normalized label matching distributions. The proposed method achieves new state-of-the-art results on all three public datasets, with a notable margin of about 3%-9% for Rank-1 accuracy compared to prior methods.

****

REVEAL: Retrieval-Augmented Visual-Language Pre-Training With Multi-Source Multimodal Knowledge Memory

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, Alireza Fathi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23369-23379

In this paper, we propose an end-to-end Retrieval-Augmented Visual Language Model (REVEAL) that learns to encode world knowledge into a large-scale memory, and to retrieve from it to answer knowledge-intensive queries. REVEAL consists of four key components: the memory, the encoder, the retriever and the generator. The large-scale memory encodes various sources of multimodal world knowledge (e.g. image-text pairs, question answering pairs, knowledge graph triplets, etc.) via a unified encoder. The retriever finds the most relevant knowledge entries in the memory, and the generator fuses the retrieved knowledge with the input query to produce the output. A key novelty in our approach is that the memory, encoder, retriever and generator are all pre-trained end-to-end on a massive amount of data. Furthermore, our approach can use a diverse set of multimodal knowledge sources, which is shown to result in significant gains. We show that REVEAL achieves state-of-the-art results on visual question answering and image captioning.

****

Learning To Retain While Acquiring: Combating Distribution-Shift in Adversarial Data-Free Knowledge Distillation

Gaurav Patel, Konda Reddy Mopuri, Qiang Qiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7786-7794

Data-free Knowledge Distillation (DFKD) has gained popularity recently, with the fundamental idea of carrying out knowledge transfer from a Teacher neural network to a Student neural network in the absence of training data. However, in the Adversarial DFKD framework, the student network's accuracy, suffers due to the non-stationary distribution of the pseudo-samples under multiple generator updates. To this end, at every generator update, we aim to maintain the student's performance on previously encountered examples while acquiring knowledge from samples of the current distribution. Thus, we propose a meta-learning inspired framework by treating the task of Knowledge-Acquisition (learning from newly generated samples) and Knowledge-Retention (retaining knowledge on previously met samples) as meta-train and meta-test, respectively. Hence, we dub our method as Learning to Retain while Acquiring. Moreover, we identify an implicit aligning factor between the Knowledge-Retention and Knowledge-Acquisition tasks indicating that the proposed student update strategy enforces a common gradient direction for both

tasks, alleviating interference between the two objectives. Finally, we support our hypothesis by exhibiting extensive evaluation and comparison of our method with prior arts on multiple datasets.
****
Why Is the Winner the Best?
Matthias Eisenmann, Annika Reinke, Vivienn Weru, Minu D. Tizabi, Fabian Isensee, Tim J. Adler, Sharib Ali, Vincent Andrearczyk, Marc Aubreville, Ujjwal Baid, Spyridon Bakas, Niranjan Balu, Sophia Bano, Jorge Bernal, Sebastian Bodenstedt, Alessandro Casella, Veronika Cheplygina, Marie Daum, Marleen de Bruijne, Adrien Depeursinge, Reuben Dorent, Jan Egger, David G. Ellis, Sandy Engelhardt, Melanie Ganz, Noha Ghatwary, Gabriel Girard, Patrick Godau, Anubha Gupta, Lasse Hansen, Kanako Harada, Mattias P. Heinrich, Nicholas Heller, Alessa Hering, Arnaud Huaulmé, Pierre Jannin, Ali Emre Kavur, Old■ich Kodym, Michal Kozubek, Jianning Li, Hongwei Li, Jun Ma, Carlos Martín-Isla, Bjoern Menze, Alison Noble, Valentin Oreiller, Nicolas Padoy, Sarthak Pati, Kelly Payette, Tim Rädsch, Jonathan Rafael-Patiño, Vivek Singh Bawa, Stefanie Speidel, Carole H. Sudre, Kimberlin van Wijnen, Martin Wagner, Donglai Wei, Amine Yamlahi, Moi Hoon Yap, Chun Yuan, Maximilian Zenk, Aneeq Zia, David Zimmerer, Dogu Baran Aydogan, Binod Bhattarai, Louise Bloch, Raphael Brüngel, Jihoon Cho, Chanyeol Choi, Qi Dou, Ivan Ezhov, Christoph M. Friedrich, Clifton D. Fuller, Rebati Raman Gaire, Adrian Galdran, Álvaro García Faura, Maria Grammatikopoulou, SeulGi Hong, Mostafa Jahanifar, Ikbeom Jang, Abdolrahim Kadkhodamohammadi, Inha Kang, Florian Kofler, Satoshi Kondo, Hugo Kuijf, Mingxing Li, Minh Luu, Tomaž Martin■i■, Pedro Morais, Mohamed A. Naser, Bruno Oliveira, David Owen, Subeen Pang, Jinah Park, Sung-Hong Park, Szymon Plotka, Elodie Puybareau, Nasir Rajpoot, Kanghyun Ryu, Numan Saeed, Adam Shephard, Pengcheng Shi, Dejan Štepec, Ronast Subedi, Guillaume Tochon, Helena R. Torres, Helene Urien, João L. Vilaça, Kareem A. Wahid, Haojie Wang, Jiacheng Wang, Liansheng Wang, Xiyue Wang, Benedikt Wiestler, Marek Wodzinski, Fangfang Xia, Juanying Xie, Zhiwei Xiong, Sen Yang, Yanwu Yang, Zixuan Zhao, Klaus Maier-Hein, Paul F. Jäger, Annette Kopp-Schneider, Lena Maier-Hein; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19955-19966
International benchmarking competitions have become fundamental for the comparative performance assessment of image analysis methods. However, little attention has been given to investigating what can be learnt from these competitions. Do they really generate scientific progress? What are common and successful participation strategies? What makes a solution superior to a competing method? To address this gap in the literature, we performed a multi-center study with all 80 competitions that were conducted in the scope of IEEE ISBI 2021 and MICCAI 2021. Statistical analyses performed based on comprehensive descriptions of the submitted algorithms linked to their rank as well as the underlying participation strategies revealed common characteristics of winning solutions. These typically include the use of multi-task learning (63%) and/or multi-stage pipelines (61%), and a focus on augmentation (100%), image preprocessing (97%), data curation (79%), and postprocessing (66%). The "typical" lead of a winning team is a computer scientist with a doctoral degree, five years of experience in biomedical image analysis, and four years of experience in deep learning. Two core general development strategies stood out for highly-ranked teams: the reflection of the metrics in the method design and the focus on analyzing and handling failure cases. According to the organizers, 43% of the winning algorithms exceeded the state of the art but only 11% completely solved the respective domain problem. The insights of our study could help researchers (1) improve algorithm development strategies when approaching new problems, and (2) focus on open research questions revealed by this work.
****
HGNet: Learning Hierarchical Geometry From Points, Edges, and Surfaces
Ting Yao, Yehao Li, Yingwei Pan, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21846-21855
Parsing an unstructured point set into constituent local geometry structures (e.g., edges or surfaces) would be helpful for understanding and representing point clouds. This motivates us to design a deep architecture to model the hierarchic

al geometry from points, edges, surfaces (triangles), to super-surfaces (adjacen
t surfaces) for the thorough analysis of point clouds. In this paper, we present
 a novel Hierarchical Geometry Network (HGNet) that integrates such hierarchical
 geometry structures from super-surfaces, surfaces, edges, to points in a top-do
wn manner for learning point cloud representations. Technically, we first constr
uct the edges between every two neighbor points. A point-level representation is
 learnt with edge-to-point aggregation, i.e., aggregating all connected edges in
to the anchor point. Next, as every two neighbor edges compose a surface, we obt
ain the edge-level representation of each anchor edge via surface-to-edge aggreg
ation over all neighbor surfaces. Furthermore, the surface-level representation
is achieved through super-surface-to-surface aggregation by transforming all sup
er-surfaces into the anchor surface. A Transformer structure is finally devised
to unify all the point-level, edge-level, and surface-level features into the ho
listic point cloud representations. Extensive experiments on four point cloud an
alysis datasets demonstrate the superiority of HGNet for 3D object classificatio
n and part/semantic segmentation tasks. More remarkably, HGNet achieves the over
all accuracy of 89.2% on ScanObjectNN, improving PointNeXt-S by 1.5%.
****
PointVector: A Vector Representation in Point Cloud Analysis
Xin Deng, WenYu Zhang, Qing Ding, XinMing Zhang; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9455-9465
In point cloud analysis, point-based methods have rapidly developed in recent ye
ars. These methods have recently focused on concise MLP structures, such as Poin
tNeXt, which have demonstrated competitiveness with Convolutional and Transforme
r structures. However, standard MLPs are limited in their ability to extract loc
al features effectively. To address this limitation, we propose a Vector-oriente
d Point Set Abstraction that can aggregate neighboring features through higher-d
imensional vectors. To facilitate network optimization, we construct a transform
ation from scalar to vector using independent angles based on 3D vector rotation
s. Finally, we develop a PointVector model that follows the structure of PointNe
Xt. Our experimental results demonstrate that PointVector achieves state-of-the-
art performance 72.3% mIOU on the S3DIS Area 5 and 78.4% mIOU on the S3DIS (6-fo
ld cross-validation) with only 58% model parameters of PointNeXt. We hope our wo
rk will help the exploration of concise and effective feature representations. T
he code will be released soon.
****
BAEFormer: Bi-Directional and Early Interaction Transformers for Bird's Eye View
 Semantic Segmentation
Cong Pan, Yonghao He, Junran Peng, Qian Zhang, Wei Sui, Zhaoxiang Zhang; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2023, pp. 9590-9599
Bird's Eye View (BEV) semantic segmentation is a critical task in autonomous dri
ving. However, existing Transformer-based methods confront difficulties in trans
forming Perspective View (PV) to BEV due to their unidirectional and posterior i
nteraction mechanisms. To address this issue, we propose a novel Bi-directional
and Early Interaction Transformers framework named BAEFormer, consisting of (i)
an early-interaction PV-BEV pipeline and (ii) a bi-directional cross-attention m
echanism. Moreover, we find that the image feature maps' resolution in the cross
-attention module has a limited effect on the final performance. Under this crit
ical observation, we propose to enlarge the size of input images and downsample
the multi-view image features for cross-interaction, further improving the accur
acy while keeping the amount of computation controllable. Our proposed method fo
r BEV semantic segmentation achieves state-of-the-art performance in real-time i
nference speed on the nuScenes dataset, i.e., 38.9 mIoU at 45 FPS on a single A1
00 GPU.
****
Good Is Bad: Causality Inspired Cloth-Debiasing for Cloth-Changing Person Re-Ide
ntification
Zhengwei Yang, Meng Lin, Xian Zhong, Yu Wu, Zheng Wang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1472

-1481
Entangled representation of clothing and identity (ID)-intrinsic clues are poten tially concomitant in conventional person Re-IDentification (ReID). Nevertheless , eliminating the negative impact of clothing on ID remains challenging due to t he lack of theory and the difficulty of isolating the exact implications. In thi s paper, a causality-based Auto-Intervention Model, referred to as AIM, is first proposed to mitigate clothing bias for robust cloth-changing person ReID (CC-Re ID). Specifically, we analyze the effect of clothing on the model inference and adopt a dual-branch model to simulate causal intervention. Progressively, clothi ng bias is eliminated automatically with model training. AIM is encouraged to le arn more discriminative ID clues that are free from clothing bias. Extensive exp eriments on two standard CC-ReID datasets demonstrate the superiority of the pro posed AIM over other state-of-the-art methods.
****
Use Your Head: Improving Long-Tail Video Recognition
Toby Perrett, Saptarshi Sinha, Tilo Burghardt, Majid Mirmehdi, Dima Damen; Proce edings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CV PR), 2023, pp. 2415-2425
This paper presents an investigation into long-tail video recognition. We demons trate that, unlike naturally-collected video datasets and existing long-tail ima ge benchmarks, current video benchmarks fall short on multiple long-tailed prope rties. Most critically, they lack few-shot classes in their tails. In response, we propose new video benchmarks that better assess long-tail recognition, by sam pling subsets from two datasets: SSv2 and VideoLT. We then propose a method, Lon g-Tail Mixed Reconstruction (LMR), which reduces overfitting to instances from f ew-shot classes by reconstructing them as weighted combinations of samples from head classes. LMR then employs label mixing to learn robust decision boundaries. It achieves state-of-the-art average class accuracy on EPIC-KITCHENS and the pr oposed SSv2-LT and VideoLT-LT. Benchmarks and code at: github.com/tobyperrett/lm r
****
Revisiting the P3P Problem
Yaqing Ding, Jian Yang, Viktor Larsson, Carl Olsson, Kalle Åström; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 3, pp. 4872-4880
One of the classical multi-view geometry problems is the so called P3P problem, where the absolute pose of a calibrated camera is determined from three 2D-to-3D correspondences. Since these solvers form a critical component of many vision s ystems (e.g. in localization and Structure-from-Motion), there have been signifi cant effort in developing faster and more stable algorithms. While the current s tate-of-the-art solvers are both extremely fast and stable, there still exist co nfigurations where they break down. In this paper we algebraically formulate the problem as finding the intersection of two conics. With this formulation we are able to analytically characterize the real roots of the polynomial system and e mploy a tailored solution strategy for each problem instance. The result is a fa st and completely stable solver, that is able to correctly solve cases where com peting methods fail. Our experimental evaluation shows that we outperform the cu rrent state-of-the-art methods both in terms of speed and success rate.
****
Generic-to-Specific Distillation of Masked Autoencoders
Wei Huang, Zhiliang Peng, Li Dong, Furu Wei, Jianbin Jiao, Qixiang Ye; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15996-16005
Large vision Transformers (ViTs) driven by self-supervised pre-training mechanis ms achieved unprecedented progress. Lightweight ViT models limited by the model capacity, however, benefit little from those pre-training mechanisms. Knowledge distillation defines a paradigm to transfer representations from large (teacher) models to small (student) ones. However, the conventional single-stage distilla tion easily gets stuck on task-specific transfer, failing to retain the task-agn ostic knowledge crucial for model generalization. In this study, we propose gene

ric-to-specific distillation (G2SD), to tap the potential of small ViT models un
der the supervision of large models pre-trained by masked autoencoders. In gener
ic distillation, decoder of the small model is encouraged to align feature predi
ctions with hidden representations of the large model, so that task-agnostic kno
wledge can be transferred. In specific distillation, predictions of the small mo
del are constrained to be consistent with those of the large model, to transfer
task-specific features which guarantee task performance. With G2SD, the vanilla
ViT-Small model respectively achieves 98.7%, 98.1% and 99.3% the performance of
its teacher (ViT-Base) for image classification, object detection, and semantic
segmentation, setting a solid baseline for two-stage vision distillation. Code w
ill be available at https://github.com/pengzhiliang/G2SD.
****
PAniC-3D: Stylized Single-View 3D Reconstruction From Portraits of Anime Charact
ers
Shuhong Chen, Kevin Zhang, Yichun Shi, Heng Wang, Yiheng Zhu, Guoxian Song, Sizh
e An, Janus Kristjansson, Xiao Yang, Matthias Zwicker; Proceedings of the IEEE/C
VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21068
-21077
We propose PAniC-3D, a system to reconstruct stylized 3D character heads directl
y from illustrated (p)ortraits of (ani)me (c)haracters. Our anime-style domain p
oses unique challenges to single-view reconstruction; compared to natural images
 of human heads, character portrait illustrations have hair and accessories with
 more complex and diverse geometry, and are shaded with non-photorealistic conto
ur lines. In addition, there is a lack of both 3D model and portrait illustratio
n data suitable to train and evaluate this ambiguous stylized reconstruction tas
k. Facing these challenges, our proposed PAniC-3D architecture crosses the illus
tration-to-3D domain gap with a line-filling model, and represents sophisticated
 geometries with a volumetric radiance field. We train our system with two large
 new datasets (11.2k Vroid 3D models, 1k Vtuber portrait illustrations), and eva
luate on a novel AnimeRecon benchmark of illustration-to-3D pairs. PAniC-3D sign
ificantly outperforms baseline methods, and provides data to establish the task
of stylized reconstruction from portrait illustrations.
****
Combining Implicit-Explicit View Correlation for Light Field Semantic Segmentati
on
Ruixuan Cong, Da Yang, Rongshan Chen, Sizhe Wang, Zhenglong Cui, Hao Sheng; Proc
eedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (C
VPR), 2023, pp. 9172-9181
Since light field simultaneously records spatial information and angular informa
tion of light rays, it is considered to be beneficial for many potential applica
tions, and semantic segmentation is one of them. The regular variation of image
information across views facilitates a comprehensive scene understanding. Howeve
r, in the case of limited memory, the high-dimensional property of light field m
akes the problem more intractable than generic semantic segmentation, manifested
 in the difficulty of fully exploiting the relationships among views while maint
aining contextual information in single view. In this paper, we propose a novel
network called LF-IENet for light field semantic segmentation. It contains two d
ifferent manners to mine complementary information from surrounding views to seg
ment central view. One is implicit feature integration that leverages attention
mechanism to compute inter-view and intra-view similarity to modulate features o
f central view. The other is explicit feature propagation that directly warps fe
atures of other views to central view under the guidance of disparity. They comp
lement each other and jointly realize complementary information fusion across vi
ews in light field. The proposed method achieves outperforming performance on bo
th real-world and synthetic light field datasets, demonstrating the effectivenes
s of this new architecture.
****
TimeBalance: Temporally-Invariant and Temporally-Distinctive Video Representatio
ns for Semi-Supervised Action Recognition
Ishan Rajendrakumar Dave, Mamshad Nayeem Rizve, Chen Chen, Mubarak Shah; Proceed

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2341-2352
Semi-Supervised Learning can be more beneficial for the video domain compared to images because of its higher annotation cost and dimensionality. Besides, any video understanding task requires reasoning over both spatial and temporal dimensions. In order to learn both the static and motion related features for the semi-supervised action recognition task, existing methods rely on hard input inductive biases like using two-modalities (RGB and Optical-flow) or two-stream of different playback rates. Instead of utilizing unlabeled videos through diverse input streams, we rely on self-supervised video representations, particularly, we utilize temporally-invariant and temporally-distinctive representations. We observe that these representations complement each other depending on the nature of the action. Based on this observation, we propose a student-teacher semi-supervised learning framework, TimeBalance, where we distill the knowledge from a temporally-invariant and a temporally-distinctive teacher. Depending on the nature of the unlabeled video, we dynamically combine the knowledge of these two teachers based on a novel temporal similarity-based reweighting scheme. Our method achieves state-of-the-art performance on three action recognition benchmarks: UCF101, HMDB51, and Kinetics400. Code: https://github.com/DAVEISHAN/TimeBalance.
****
RiDDLE: Reversible and Diversified De-Identification With Latent Encryptor
Dongze Li, Wei Wang, Kang Zhao, Jing Dong, Tieniu Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8093-8102
This work presents RiDDLE, short for Reversible and Diversified De-identification with Latent Encryptor, to protect the identity information of people from being misused. Built upon a pre-learned StyleGAN2 generator, RiDDLE manages to encrypt and decrypt the facial identity within the latent space. The design of RiDDLE has three appealing properties. First, the encryption process is cipher-guided and hence allows diverse anonymization using different passwords. Second, the true identity can only be decrypted with the correct password, otherwise the system will produce another de-identified face to maintain the privacy. Third, both encryption and decryption share an efficient implementation, benefiting from a carefully tailored lightweight encryptor. Comparisons with existing alternatives confirm that our approach accomplishes the de-identification task with better quality, higher diversity, and stronger reversibility. We further demonstrate the effectiveness of RiDDLE in anonymizing videos. Code and models will be made publicly available.
****
SunStage: Portrait Reconstruction and Relighting Using the Sun as a Light Stage
Yifan Wang, Aleksander Holynski, Xiuming Zhang, Xuaner Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20792-20802
A light stage uses a series of calibrated cameras and lights to capture a subject's facial appearance under varying illumination and viewpoint. This captured information is crucial for facial reconstruction and relighting. Unfortunately, light stages are often inaccessible: they are expensive and require significant technical expertise for construction and operation. In this paper, we present SunStage: a lightweight alternative to a light stage that captures comparable data using only a smartphone camera and the sun. Our method only requires the user to capture a selfie video outdoors, rotating in place, and uses the varying angles between the sun and the face as guidance in joint reconstruction of facial geometry, reflectance, camera pose, and lighting parameters. Despite the in-the-wild un-calibrated setting, our approach is able to reconstruct detailed facial appearance and geometry, enabling compelling effects such as relighting, novel view synthesis, and reflectance editing.
****
Private Image Generation With Dual-Purpose Auxiliary Classifier
Chen Chen, Daochang Liu, Siqi Ma, Surya Nepal, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 203

61-20370
Privacy-preserving image generation has been important for segments such as medi cal domains that have sensitive and limited data. The benefits of guaranteed pri vacy come at the costs of generated images' quality and utility due to the priva cy budget constraints. The utility is currently measured by the gen2real accurac y (g2r%), i.e., the accuracy on real data of a downstream classifier trained usi ng generated data. However, apart from this standard utility, we identify the "r eversed utility" as another crucial aspect, which computes the accuracy on gener ated data of a classifier trained using real data, dubbed as real2gen accuracy (r2g%). Jointly considering these two views of utility, the standard and the reve rsed, could help the generation model better improve transferability between fak e and real data. Therefore, we propose a novel private image generation method t hat incorporates a dual-purpose auxiliary classifier, which alternates between l earning from real data and fake data, into the training of differentially privat e GANs. Additionally, our deliberate training strategies such as sequential trai ning contributes to accelerating the generator's convergence and further boostin g the performance upon exhausting the privacy budget. Our results achieve new st ate-of-the-arts over all metrics on three benchmarks: MNIST, Fashion-MNIST, and CelebA.
****
3D-POP - An Automated Annotation Approach to Facilitate Markerless 2D-3D Trackin g of Freely Moving Birds With Marker-Based Motion Capture
Hemal Naik, Alex Hoi Hang Chan, Junran Yang, Mathilde Delacoux, Iain D. Couzin, Fumihiro Kano, Máté Nagy; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2023, pp. 21274-21284
Recent advances in machine learning and computer vision are revolutionizing the field of animal behavior by enabling researchers to track the poses and location s of freely moving animals without any marker attachment. However, large dataset s of annotated images of animals for markerless pose tracking, especially high-r esolution images taken from multiple angles with accurate 3D annotations, are st ill scant. Here, we propose a method that uses a motion capture (mo-cap) system to obtain a large amount of annotated data on animal movement and posture (2D an d 3D) in a semi-automatic manner. Our method is novel in that it extracts the 3D positions of morphological keypoints (e.g eyes, beak, tail) in reference to the positions of markers attached to the animals. Using this method, we obtained, a nd offer here, a new dataset - 3D-POP with approximately 300k annotated frames (4 million instances) in the form of videos having groups of one to ten freely mo ving birds from 4 different camera views in a 3.6m x 4.2m area. 3D-POP is the fi rst dataset of flocking birds with accurate keypoint annotations in 2D and 3D al ong with bounding box and individual identities and will facilitate the developm ent of solutions for problems of 2D to 3D markerless pose, trajectory tracking, and identification in birds.
****
SOOD: Towards Semi-Supervised Oriented Object Detection
Wei Hua, Dingkang Liang, Jingyu Li, Xiaolong Liu, Zhikang Zou, Xiaoqing Ye, Xian g Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Rec ognition (CVPR), 2023, pp. 15558-15567
Semi-Supervised Object Detection (SSOD), aiming to explore unlabeled data for bo osting object detectors, has become an active task in recent years. However, exi sting SSOD approaches mainly focus on horizontal objects, leaving multi-oriented objects that are common in aerial images unexplored. This paper proposes a nove l Semi-supervised Oriented Object Detection model, termed SOOD, built upon the m ainstream pseudo-labeling framework. Towards oriented objects in aerial scenes, we design two loss functions to provide better supervision. Focusing on the orie ntations of objects, the first loss regularizes the consistency between each pse udo-label-prediction pair (includes a prediction and its corresponding pseudo la bel) with adaptive weights based on their orientation gap. Focusing on the layou t of an image, the second loss regularizes the similarity and explicitly builds the many-to-many relation between the sets of pseudo-labels and predictions. Suc h a global consistency constraint can further boost semi-supervised learning. Ou

r experiments show that when trained with the two proposed losses, SOOD surpasses the state-of-the-art SSOD methods under various settings on the DOTA-v1.5 benchmark. The code will be available at https://github.com/HamPerdredes/SOOD.
****
Unified Keypoint-Based Action Recognition Framework via Structured Keypoint Pooling
Ryo Hachiuma, Fumiaki Sato, Taiki Sekii; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22962-22971
This paper simultaneously addresses three limitations associated with conventional skeleton-based action recognition; skeleton detection and tracking errors, poor variety of the targeted actions, as well as person-wise and frame-wise action recognition. A point cloud deep-learning paradigm is introduced to the action recognition, and a unified framework along with a novel deep neural network architecture called Structured Keypoint Pooling is proposed. The proposed method sparsely aggregates keypoint features in a cascaded manner based on prior knowledge of the data structure (which is inherent in skeletons), such as the instances and frames to which each keypoint belongs, and achieves robustness against input errors. Its less constrained and tracking-free architecture enables time-series keypoints consisting of human skeletons and nonhuman object contours to be efficiently treated as an input 3D point cloud and extends the variety of the targeted action. Furthermore, we propose a Pooling-Switching Trick inspired by Structured Keypoint Pooling. This trick switches the pooling kernels between the training and inference phases to detect person-wise and frame-wise actions in a weakly supervised manner using only video-level action labels. This trick enables our training scheme to naturally introduce novel data augmentation, which mixes multiple point clouds extracted from different videos. In the experiments, we comprehensively verify the effectiveness of the proposed method against the limitations, and the method outperforms state-of-the-art skeleton-based action recognition and spatio-temporal action localization methods.
****
Multi-View Reconstruction Using Signed Ray Distance Functions (SRDF)
Pierre Zins, Yuanlu Xu, Edmond Boyer, Stefanie Wuhrer, Tony Tung; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16696-16706
In this paper, we investigate a new optimization framework for multi-view 3D shape reconstructions. Recent differentiable rendering approaches have provided breakthrough performances with implicit shape representations though they can still lack precision in the estimated geometries. On the other hand multi-view stereo methods can yield pixel wise geometric accuracy with local depth predictions along viewing rays. Our approach bridges the gap between the two strategies with a novel volumetric shape representation that is implicit but parameterized with pixel depths to better materialize the shape surface with consistent signed distances along viewing rays. The approach retains pixel-accuracy while benefiting from volumetric integration in the optimization. To this aim, depths are optimized by evaluating, at each 3D location within the volumetric discretization, the agreement between the depth prediction consistency and the photometric consistency for the corresponding pixels. The optimization is agnostic to the associated photo-consistency term which can vary from a median-based baseline to more elaborate criteria, learned functions. Our experiments demonstrate the benefit of the volumetric integration with depth predictions. They also show that our approach outperforms existing approaches over standard 3D benchmarks with better geometry estimations.
****
Beyond mAP: Towards Better Evaluation of Instance Segmentation
Rohit Jena, Lukas Zhornyak, Nehal Doiphode, Pratik Chaudhari, Vivek Buch, James Gee, Jianbo Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11309-11318
Correctness of instance segmentation constitutes counting the number of objects, correctly localizing all predictions and classifying each localized prediction. Average Precision is the de-facto metric used to measure all these constituents

 of segmentation. However, this metric does not penalize duplicate predictions i
n the high-recall range, and cannot distinguish instances that are localized cor
rectly but categorized incorrectly. This weakness has inadvertently led to netwo
rk designs that achieve significant gains in AP but also introduce a large numbe
r of false positives. We therefore cannot rely on AP to choose a model that prov
ides an optimal tradeoff between false positives and high recall. To resolve thi
s dilemma, we review alternative metrics in the literature and propose two new m
easures to explicitly measure the amount of both spatial and categorical duplica
te predictions. We also propose a Semantic Sorting and NMS module to remove thes
e duplicates based on a pixel occupancy matching scheme. Experiments show that m
odern segmentation networks have significant gains in AP, but also contain a con
siderable amount of duplicates. Our Semantic Sorting and NMS can be added as a p
lug-and-play module to mitigate hedged predictions and preserve AP.
****
Generating Aligned Pseudo-Supervision From Non-Aligned Data for Image Restoratio
n in Under-Display Camera
Ruicheng Feng, Chongyi Li, Huaijin Chen, Shuai Li, Jinwei Gu, Chen Change Loy; P
roceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
 (CVPR), 2023, pp. 5013-5022
Due to the difficulty in collecting large-scale and perfectly aligned paired tra
ining data for Under-Display Camera (UDC) image restoration, previous methods re
sort to monitor-based image systems or simulation-based methods, sacrificing the
 realness of the data and introducing domain gaps. In this work, we revisit the
classic stereo setup for training data collection -- capturing two images of the
 same scene with one UDC and one standard camera. The key idea is to "copy" deta
ils from a high-quality reference image and "paste" them on the UDC image. While
 being able to generate real training pairs, this setting is susceptible to spat
ial misalignment due to perspective and depth of field changes. The problem is f
urther compounded by the large domain discrepancy between the UDC and normal ima
ges, which is unique to UDC restoration. In this paper, we mitigate the non-triv
ial domain discrepancy and spatial misalignment through a novel Transformer-base
d framework that generates well-aligned yet high-quality target data for the cor
responding UDC input. This is made possible through two carefully designed compo
nents, namely, the Domain Alignment Module (DAM) and Geometric Alignment Module
(GAM), which encourage robust and accurate discovery of correspondence between t
he UDC and normal views. Extensive experiments show that high-quality and well-a
ligned pseudo UDC training pairs are beneficial for training a robust restoratio
n network. Code and the dataset are available at https://github.com/jnjaby/Align
Former.
****
Improving Cross-Modal Retrieval With Set of Diverse Embeddings
Dongwon Kim, Namyup Kim, Suha Kwak; Proceedings of the IEEE/CVF Conference on Co
mputer Vision and Pattern Recognition (CVPR), 2023, pp. 23422-23431
Cross-modal retrieval across image and text modalities is a challenging task due
 to its inherent ambiguity: An image often exhibits various situations, and a ca
ption can be coupled with diverse images. Set-based embedding has been studied a
s a solution to this problem. It seeks to encode a sample into a set of differen
t embedding vectors that capture different semantics of the sample. In this pape
r, we present a novel set-based embedding method, which is distinct from previou
s work in two aspects. First, we present a new similarity function called smooth
-Chamfer similarity, which is designed to alleviate the side effects of existing
 similarity functions for set-based embedding. Second, we propose a novel set pr
ediction module to produce a set of embedding vectors that effectively captures
diverse semantics of input by the slot attention mechanism. Our method is evalua
ted on the COCO and Flickr30K datasets across different visual backbones, where
it outperforms existing methods including ones that demand substantially larger
computation at inference.
****
BASiS: Batch Aligned Spectral Embedding Space
Or Streicher, Ido Cohen, Guy Gilboa; Proceedings of the IEEE/CVF Conference on C

omputer Vision and Pattern Recognition (CVPR), 2023, pp. 10396-10405
Graph is a highly generic and diverse representation, suitable for almost any da ta processing problem. Spectral graph theory has been shown to provide powerful algorithms, backed by solid linear algebra theory. It thus can be extremely inst rumental to design deep network building blocks with spectral graph characterist ics. For instance, such a network allows the design of optimal graphs for certai n tasks or obtaining a canonical orthogonal low-dimensional embedding of the dat a. Recent attempts to solve this problem were based on minimizing Rayleigh-quoti ent type losses. We propose a different approach of directly learning the graph' s eigensapce. A severe problem of the direct approach, applied in batch-learning , is the inconsistent mapping of features to eigenspace coordinates in different batches. We analyze the degrees of freedom of learning this task using batches and propose a stable alignment mechanism that can work both with batch changes a nd with graph-metric changes. We show that our learnt spectral embedding is bett er in terms of NMI, ACC, Grassman distnace, orthogonality and classification acc uracy, compared to SOTA. In addition, the learning is more stable.
****
Neural Pixel Composition for 3D-4D View Synthesis From Multi-Views
Aayush Bansal, Michael Zollhöfer; Proceedings of the IEEE/CVF Conference on Comp uter Vision and Pattern Recognition (CVPR), 2023, pp. 290-299
We present Neural Pixel Composition (NPC), a novel approach for continuous 3D-4D view synthesis given only a discrete set of multi-view observations as input. E xisting state-of-the-art approaches require dense multi-view supervision and an extensive computational budget. The proposed formulation reliably operates on sp arse and wide-baseline multi-view imagery and can be trained efficiently within a few seconds to 10 minutes for hi-res (12MP) content, i.e., 200-400X faster con vergence than existing methods. Crucial to our approach are two core novelties: 1) a representation of a pixel that contains color and depth information accumul ated from multi-views for a particular location and time along a line of sight, and 2) a multi-layer perceptron (MLP) that enables the composition of this rich information provided for a pixel location to obtain the final color output. We e xperiment with a large variety of multi-view sequences, compare to existing appr oaches, and achieve better results in diverse and challenging settings.
****
DCFace: Synthetic Face Generation With Dual Condition Diffusion Model
Minchul Kim, Feng Liu, Anil Jain, Xiaoming Liu; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12715-12725
Generating synthetic datasets for training face recognition models is challengin g because dataset generation entails more than creating high fidelity images. It involves generating multiple images of same subjects under different factors (e .g., variations in pose, illumination, expression, aging and occlusion) which fo llows the real image conditional distribution. Previous works have studied the g eneration of synthetic datasets using GAN or 3D models. In this work, we approac h the problem from the aspect of combining subject appearance (ID) and external factor (style) conditions. These two conditions provide a direct way to control the inter-class and intra-class variations. To this end, we propose a Dual Condi tion Face Generator (DCFace) based on a diffusion model. Our novel Patch-wise st yle extractor and Time-step dependent ID loss enables DCFace to consistently pro duce face images of the same subject under different styles with precise control . Face recognition models trained on synthetic images from the proposed DCFace p rovide higher verification accuracies compared to previous works by 6.11% on ave rage in 4 out of 5 test datasets, LFW, CFP-FP, CPLFW, AgeDB and CALFW. Model, co de, and synthetic dataset are available at https://github.com/mk-minchul/dcface
****
CRAFT: Concept Recursive Activation FacTorization for Explainability
Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Jul ien Colin, Rémi Cadène, Thomas Serre; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2711-2721
Attribution methods are a popular class of explainability methods that use heatm aps to depict the most important areas of an image that drive a model decision.

Nevertheless, recent work has shown that these methods have limited utility in practice, presumably because they only highlight the most salient parts of an image (i.e., "where" the model looked) and do not communicate any information about "what" the model saw at those locations. In this work, we try to fill in this gap with Craft -- a novel approach to identify both "what" and "where" by generating concept-based explanations. We introduce 3 new ingredients to the automatic concept extraction literature: (i) a recursive strategy to detect and decompose concepts across layers, (ii) a novel method for a more faithful estimation of concept importance using Sobol indices, and (iii) the use of implicit differentiation to unlock Concept Attribution Maps. We conduct both human and computer vision experiments to demonstrate the benefits of the proposed approach. We show that our recursive decomposition generates meaningful and accurate concepts and that the proposed concept importance estimation technique is more faithful to the model than previous methods. When evaluating the usefulness of the method for human experimenters on the utility benchmark, we find that our approach significantly improves on two of the three test scenarios (while none of the current methods including ours help on the third). Overall, our study suggests that, while much work remains toward the development of general explainability methods that are useful in practical scenarios, the identification of meaningful concepts at the proper level of granularity yields useful and complementary information beyond that afforded by attribution methods.
****
Policy Adaptation From Foundation Model Feedback
Yuying Ge, Annabella Macaluso, Li Erran Li, Ping Luo, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19059-19069
Recent progress on vision-language foundation models have brought significant advancement to building general-purpose robots. By using the pre-trained models to encode the scene and instructions as inputs for decision making, the instruction-conditioned policy can generalize across different objects and tasks. While this is encouraging, the policy still fails in most cases given an unseen task or environment. In this work, we propose Policy Adaptation from Foundation model Feedback (PAFF). When deploying the trained policy to a new task or a new environment, we first let the policy play with randomly generated instructions to record the demonstrations. While the execution could be wrong, we can use the pre-trained foundation models to provide feedback to relabel the demonstrations. This automatically provides new pairs of demonstration-instruction data for policy fine-tuning. We evaluate our method on a broad range of experiments with the focus on generalization on unseen objects, unseen tasks, unseen environments, and sim-to-real transfer. We show PAFF improves baselines by a large margin in all cases.
****
Recognizing Rigid Patterns of Unlabeled Point Clouds by Complete and Continuous Isometry Invariants With No False Negatives and No False Positives
Daniel Widdowson, Vitaliy Kurlin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1275-1284
Rigid structures such as cars or any other solid objects are often represented by finite clouds of unlabeled points. The most natural equivalence on these point clouds is rigid motion or isometry maintaining all inter-point distances. Rigid patterns of point clouds can be reliably compared only by complete isometry invariants that can also be called equivariant descriptors without false negatives (isometric clouds having different descriptions) and without false positives (non-isometric clouds with the same description). Noise and motion in data motivate a search for invariants that are continuous under perturbations of points in a suitable metric. We propose the first continuous and complete invariant of unlabeled clouds in any Euclidean space. For a fixed dimension, the new metric for this invariant is computable in a polynomial time in the number of points.
****
N-Gram in Swin Transformers for Efficient Lightweight Image Super-Resolution
Haram Choi, Jeongmin Lee, Jihoon Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2071-2081

While some studies have proven that Swin Transformer (Swin) with window self-att ention (WSA) is suitable for single image super-resolution (SR), the plain WSA i gnores the broad regions when reconstructing high-resolution images due to a lim ited receptive field. In addition, many deep learning SR methods suffer from int ensive computations. To address these problems, we introduce the N-Gram context to the low-level vision with Transformers for the first time. We define N-Gram a s neighboring local windows in Swin, which differs from text analysis that views N-Gram as consecutive characters or words. N-Grams interact with each other by sliding-WSA, expanding the regions seen to restore degraded pixels. Using the N-Gram context, we propose NGswin, an efficient SR network with SCDP bottleneck ta king multi-scale outputs of the hierarchical encoder. Experimental results show that NGswin achieves competitive performance while maintaining an efficient stru cture when compared with previous leading methods. Moreover, we also improve oth er Swin-based SR methods with the N-Gram context, thereby building an enhanced m odel: SwinIR-NG. Our improved SwinIR-NG outperforms the current best lightweight SR approaches and establishes state-of-the-art results. Codes are available at https://github.com/rami0205/NGramSwin.
****

Semi-DETR: Semi-Supervised Object Detection With Detection Transformers

Jiacheng Zhang, Xiangru Lin, Wei Zhang, Kuo Wang, Xiao Tan, Junyu Han, Errui Din g, Jingdong Wang, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23809-23818

We analyze the DETR-based framework on semi-supervised object detection (SSOD) a nd observe that (1) the one-to-one assignment strategy generates incorrect match ing when the pseudo ground-truth bounding box is inaccurate, leading to training inefficiency; (2) DETR-based detectors lack deterministic correspondence betwee n the input query and its prediction output, which hinders the applicability of the consistency-based regularization widely used in current SSOD methods. We pre sent Semi-DETR, the first transformer-based end-to-end semi-supervised object de tector, to tackle these problems. Specifically, we propose a Stage-wise Hybrid M atching strategy that com- bines the one-to-many assignment and one-to-one assig nment strategies to improve the training efficiency of the first stage and thus provide high-quality pseudo labels for the training of the second stage. Besides , we introduce a Cross-view Query Consistency method to learn the semantic featu re invariance of object queries from different views while avoiding the need to find deterministic query correspondence. Furthermore, we propose a Cost-based Ps eudo Label Mining module to dynamically mine more pseudo boxes based on the matc hing cost of pseudo ground truth bounding boxes for consistency training. Extens ive experiments on all SSOD settings of both COCO and Pascal VOC benchmark datas ets show that our Semi-DETR method outperforms all state-of-the-art methods by c lear margins.
****

Infinite Photorealistic Worlds Using Procedural Generation

Alexander Raistrick, Lahav Lipson, Zeyu Ma, Lingjie Mei, Mingzhe Wang, Yiming Zu o, Karhan Kayan, Hongyu Wen, Beining Han, Yihan Wang, Alejandro Newell, Hei Law, Ankit Goyal, Kaiyu Yang, Jia Deng; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2023, pp. 12630-12641

We introduce Infinigen, a procedural generator of photorealistic 3D scenes of th e natural world. Infinigen is entirely procedural: every asset, from shape to te xture, is generated from scratch via randomized mathematical rules, using no ext ernal source and allowing infinite variation and composition. Infinigen offers b road coverage of objects and scenes in the natural world including plants, anima ls, terrains, and natural phenomena such as fire, cloud, rain, and snow. Infinig en can be used to generate unlimited, diverse training data for a wide range of computer vision tasks including object detection, semantic segmentation, optical flow, and 3D reconstruction. We expect Infinigen to be a useful resource for co mputer vision research and beyond. Please visit https://infinigen.org for videos , code and pre-generated data.
****

Diversity-Measurable Anomaly Detection

Wenrui Liu, Hong Chang, Bingpeng Ma, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12147-12156

Reconstruction-based anomaly detection models achieve their purpose by suppressing the generalization ability for anomaly. However, diverse normal patterns are consequently not well reconstructed as well. Although some efforts have been made to alleviate this problem by modeling sample diversity, they suffer from shortcut learning due to undesired transmission of abnormal information. In this paper, to better solve the tradeoff problem, we propose Diversity-Measurable Anomaly Detection (DMAD) framework to enhance reconstruction diversity while avoid the undesired generalization on anomalies. To this end, we design Pyramid Deformation Module (PDM), which models diverse normals and measures the severity of anomaly by estimating multi-scale deformation fields from reconstructed reference to original input. Integrated with an information compression module, PDM essentially decouples deformation from prototypical embedding and makes the final anomaly score more reliable. Experimental results on both surveillance videos and industrial images demonstrate the effectiveness of our method. In addition, DMAD works equally well in front of contaminated data and anomaly-like normal samples.

****

Hybrid Neural Rendering for Large-Scale Scenes With Motion Blur

Peng Dai, Yinda Zhang, Xin Yu, Xiaoyang Lyu, Xiaojuan Qi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 154-164

Rendering novel view images is highly desirable for many applications. Despite recent progress, it remains challenging to render high-fidelity and view-consistent novel views of large-scale scenes from in-the-wild images with inevitable artifacts (e.g., motion blur). To this end, we develop a hybrid neural rendering model that makes image-based representation and neural 3D representation join forces to render high-quality, view-consistent images. Besides, images captured in the wild inevitably contain artifacts, such as motion blur, which deteriorates the quality of rendered images. Accordingly, we propose strategies to simulate blur effects on the rendered images to mitigate the negative influence of blurriness images and reduce their importance during training based on precomputed quality-aware weights. Extensive experiments on real and synthetic data demonstrate our model surpasses state-of-the-art point-based methods for novel view synthesis. The code is available at https://daipengwa.github.io/Hybrid-Rendering-ProjectPage.

****

Perception-Oriented Single Image Super-Resolution Using Optimal Objective Estimation

Seung Ho Park, Young Su Moon, Nam Ik Cho; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1725-1735

Single-image super-resolution (SISR) networks trained with perceptual and adversarial losses provide high-contrast outputs compared to those of networks trained with distortion-oriented losses, such as L1 or L2. However, it has been shown that using a single perceptual loss is insufficient for accurately restoring locally varying diverse shapes in images, often generating undesirable artifacts or unnatural details. For this reason, combinations of various losses, such as perceptual, adversarial, and distortion losses, have been attempted, yet it remains challenging to find optimal combinations. Hence, in this paper, we propose a new SISR framework that applies optimal objectives for each region to generate plausible results in overall areas of high-resolution outputs. Specifically, the framework comprises two models: a predictive model that infers an optimal objective map for a given low-resolution (LR) input and a generative model that applies a target objective map to produce the corresponding SR output. The generative model is trained over our proposed objective trajectory representing a set of essential objectives, which enables the single network to learn various SR results corresponding to combined losses on the trajectory. The predictive model is trained using pairs of LR images and corresponding optimal objective maps searched from the objective trajectory. Experimental results on five benchmarks show that th

e proposed method outperforms state-of-the-art perception-driven SR methods in LPIPS, DISTS, PSNR, and SSIM metrics. The visual results also demonstrate the superiority of our method in perception-oriented reconstruction. The code is available at https://github.com/seungho-snu/SROOE.
****

GP-VTON: Towards General Purpose Virtual Try-On via Collaborative Local-Flow Global-Parsing Learning

Zhenyu Xie, Zaiyu Huang, Xin Dong, Fuwei Zhao, Haoye Dong, Xijin Zhang, Feida Zhu, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23550-23559

Image-based Virtual Try-ON aims to transfer an in-shop garment onto a specific person. Existing methods employ a global warping module to model the anisotropic deformation for different garment parts, which fails to preserve the semantic information of different parts when receiving challenging inputs (e.g, intricate human poses, difficult garments). Moreover, most of them directly warp the input garment to align with the boundary of the preserved region, which usually requires texture squeezing to meet the boundary shape constraint and thus leads to texture distortion. The above inferior performance hinders existing methods from real-world applications. To address these problems and take a step towards real-world virtual try-on, we propose a General-Purpose Virtual Try-ON framework, named GP-VTON, by developing an innovative Local-Flow Global-Parsing (LFGP) warping module and a Dynamic Gradient Truncation (DGT) training strategy. Specifically, compared with the previous global warping mechanism, LFGP employs local flows to warp garments parts individually, and assembles the local warped results via the global garment parsing, resulting in reasonable warped parts and a semantic-correct intact garment even with challenging inputs.On the other hand, our DGT training strategy dynamically truncates the gradient in the overlap area and the warped garment is no more required to meet the boundary constraint, which effectively avoids the texture squeezing problem. Furthermore, our GP-VTON can be easily extended to multi-category scenario and jointly trained by using data from different garment categories. Extensive experiments on two high-resolution benchmarks demonstrate our superiority over the existing state-of-the-art methods.
****

A Large-Scale Robustness Analysis of Video Action Recognition Models

Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, Yogesh S. Rawat; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14698-14708

We have seen great progress in video action recognition in recent years. There are several models based on convolutional neural network (CNN) and some recent transformer based approaches which provide top performance on existing benchmarks. In this work, we perform a large-scale robustness analysis of these existing models for video action recognition. We focus on robustness against real-world distribution shift perturbations instead of adversarial perturbations. We propose four different benchmark datasets, HMDB51-P, UCF101-P, Kinetics400-P, and SSv2-P to perform this analysis. We study robustness of six state-of-the-art action recognition models against 90 different perturbations. The study reveals some interesting findings, 1) Transformer based models are consistently more robust compared to CNN based models, 2) Pre-training improves robustness for Transformer based models more than CNN based models, and 3) All of the studied models are robust to temporal perturbations for all datasets but SSv2; suggesting the importance of temporal information for action recognition varies based on the dataset and activities. Next, we study the role of augmentations in model robustness and present a real-world dataset, UCF101-DS, which contains realistic distribution shifts, to further validate some of these findings. We believe this study will serve as a benchmark for future research in robust video action recognition.
****

Decomposed Soft Prompt Guided Fusion Enhancing for Compositional Zero-Shot Learning

Xiaocheng Lu, Song Guo, Ziming Liu, Jingcai Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23560-23569

Compositional Zero-Shot Learning (CZSL) aims to recognize novel concepts formed by known states and objects during training. Existing methods either learn the combined state-object representation, challenging the generalization of unseen compositions, or design two classifiers to identify state and object separately from image features, ignoring the intrinsic relationship between them. To jointly eliminate the above issues and construct a more robust CZSL system, we propose a novel framework termed Decomposed Fusion with Soft Prompt (DFSP), by involving vision-language models (VLMs) for unseen composition recognition. Specifically, DFSP constructs a vector combination of learnable soft prompts with state and object to establish the joint representation of them. In addition, a cross-modal decomposed fusion module is designed between the language and image branches, which decomposes state and object among language features instead of image features. Notably, being fused with the decomposed features, the image features can be more expressive for learning the relationship with states and objects, respectively, to improve the response of unseen compositions in the pair space, hence narrowing the domain gap between seen and unseen sets. Experimental results on three challenging benchmarks demonstrate that our approach significantly outperforms other state-of-the-art methods by large margins.
****
Hierarchical Semantic Contrast for Scene-Aware Video Anomaly Detection
Shengyang Sun, Xiaojin Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22846-22856
Increasing scene-awareness is a key challenge in video anomaly detection (VAD). In this work, we propose a hierarchical semantic contrast (HSC) method to learn a scene-aware VAD model from normal videos. We first incorporate foreground object and background scene features with high-level semantics by taking advantage of pre-trained video parsing models. Then, building upon the autoencoder-based reconstruction framework, we introduce both scene-level and object-level contrastive learning to enforce the encoded latent features to be compact within the same semantic classes while being separable across different classes. This hierarchical semantic contrast strategy helps to deal with the diversity of normal patterns and also increases their discrimination ability. Moreover, for the sake of tackling rare normal activities, we design a skeleton-based motion augmentation to increase samples and refine the model further. Extensive experiments on three public datasets and scene-dependent mixture datasets validate the effectiveness of our proposed method.
****
All-in-Focus Imaging From Event Focal Stack
Hanyue Lou, Minggui Teng, Yixin Yang, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17366-17375
Traditional focal stack methods require multiple shots to capture images focused at different distances of the same scene, which cannot be applied to dynamic scenes well. Generating a high-quality all-in-focus image from a single shot is challenging, due to the highly ill-posed nature of the single-image defocus and deblurring problem. In this paper, to restore an all-in-focus image, we propose the event focal stack which is defined as event streams captured during a continuous focal sweep. Given an RGB image focused at an arbitrary distance, we explore the high temporal resolution of event streams, from which we automatically select refocusing timestamps and reconstruct corresponding refocused images with events to form a focal stack. Guided by the neighbouring events around the selected timestamps, we can merge the focal stack with proper weights and restore a sharp all-in-focus image. Experimental results on both synthetic and real datasets show superior performance over state-of-the-art methods.
****
Video Probabilistic Diffusion Models in Projected Latent Space
Sihyun Yu, Kihyuk Sohn, Subin Kim, Jinwoo Shin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18456-18466
Despite the remarkable progress in deep generative models, synthesizing high-resolution and temporally coherent videos still remains a challenge due to their high-dimensionality and complex temporal dynamics along with large spatial variati

ons. Recent works on diffusion models have shown their potential to solve this c hallenge, yet they suffer from severe computation- and memory-inefficiency that limit the scalability. To handle this issue, we propose a novel generative model for videos, coined projected latent video diffusion models (PVDM), a probabilis tic diffusion model which learns a video distribution in a low-dimensional laten t space and thus can be efficiently trained with high-resolution videos under li mited resources. Specifically, PVDM is composed of two components: (a) an autoen coder that projects a given video as 2D-shaped latent vectors that factorize the complex cubic structure of video pixels and (b) a diffusion model architecture specialized for our new factorized latent space and the training/sampling proced ure to synthesize videos of arbitrary length with a single model. Experiments on popular video generation datasets demonstrate the superiority of PVDM compared with previous video synthesis methods; e.g., PVDM obtains the FVD score of 639.7 on the UCF-101 long video (128 frames) generation benchmark, which improves 177 3.4 of the prior state-of-the-art.
****
Learning 3D Scene Priors With 2D Supervision
Yinyu Nie, Angela Dai, Xiaoguang Han, Matthias Nießner; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 792-802
Holistic 3D scene understanding entails estimation of both layout configuration and object geometry in a 3D environment. Recent works have shown advances in 3D scene estimation from various input modalities (e.g., images, 3D scans), by leve raging 3D supervision (e.g., 3D bounding boxes or CAD models), for which collect ion at scale is expensive and often intractable. To address this shortcoming, we propose a new method to learn 3D scene priors of layout and shape without requi ring any 3D ground truth. Instead, we rely on 2D supervision from multi-view RGB images. Our method represents a 3D scene as a latent vector, from which we can progressively decode to a sequence of objects characterized by their class categ ories, 3D bounding boxes, and meshes. With our trained autoregressive decoder re presenting the scene prior, our method facilitates many downstream applications, including scene synthesis, interpolation, and single-view reconstruction. Exper iments on 3D-FRONT and ScanNet show that our method outperforms state of the art in single-view reconstruction, and achieves state-of-the-art results in scene s ynthesis against baselines which require for 3D supervision.
****
Blind Video Deflickering by Neural Filtering With a Flawed Atlas
Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, Qifeng Chen; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 104 39-10448
Many videos contain flickering artifacts; common causes of flicker include video processing algorithms, video generation algorithms, and capturing videos under specific situations. Prior work usually requires specific guidance such as the f lickering frequency, manual annotations, or extra consistent videos to remove th e flicker. In this work, we propose a general flicker removal framework that onl y receives a single flickering video as input without additional guidance. Since it is blind to a specific flickering type or guidance, we name this "blind defl ickering." The core of our approach is utilizing the neural atlas in cooperation with a neural filtering strategy. The neural atlas is a unified representation for all frames in a video that provides temporal consistency guidance but is fla wed in many cases. To this end, a neural network is trained to mimic a filter to learn the consistent features (e.g., color, brightness) and avoid introducing t he artifacts in the atlas. To validate our method, we construct a dataset that c ontains diverse real-world flickering videos. Extensive experiments show that ou r method achieves satisfying deflickering performance and even outperforms basel ines that use extra guidance on a public benchmark. The source code is publicly available at https://chenyanglei.github.io/deflicker.
****
Label-Free Liver Tumor Segmentation
Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L. Yuille, Z

ongwei Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte
rn Recognition (CVPR), 2023, pp. 7422-7432
We demonstrate that AI models can accurately segment liver tumors without the need for manual annotation by using synthetic tumors in CT scans. Our synthetic tumors have two intriguing advantages: (I) realistic in shape and texture, which even medical professionals can confuse with real tumors; (II) effective for training AI models, which can perform liver tumor segmentation similarly to the model trained on real tumors--this result is exciting because no existing work, using synthetic tumors only, has thus far reached a similar or even close performance to real tumors. This result also implies that manual efforts for annotating tumors voxel by voxel (which took years to create) can be significantly reduced in the future. Moreover, our synthetic tumors can automatically generate many examples of small (or even tiny) synthetic tumors and have the potential to improve the success rate of detecting small liver tumors, which is critical for detecting the early stages of cancer. In addition to enriching the training data, our synthesizing strategy also enables us to rigorously assess the AI robustness.
****
Grid-Guided Neural Radiance Fields for Large Urban Scenes
Linning Xu, Yuanbo Xiangli, Sida Peng, Xingang Pan, Nanxuan Zhao, Christian Theobalt, Bo Dai, Dahua Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8296-8306
Purely MLP-based neural radiance fields (NeRF-based methods) often suffer from underfitting with blurred renderings on large-scale scenes due to limited model capacity. Recent approaches propose to geographically divide the scene and adopt multiple sub-NeRFs to model each region individually, leading to linear scale-up in training costs and the number of sub-NeRFs as the scene expands. An alternative solution is to use a feature grid representation, which is computationally efficient and can naturally scale to a large scene with increased grid resolutions. However, the feature grid tends to be less constrained and often reaches suboptimal solutions, producing noisy artifacts in renderings, especially in regions with complex geometry and texture. In this work, we present a new framework that realizes high-fidelity rendering on large urban scenes while being computationally efficient. We propose to use a compact multi-resolution ground feature plane representation to coarsely capture the scene, and complement it with positional encoding inputs through another NeRF branch for rendering in a joint learning fashion. We show that such an integration can utilize the advantages of two alternative solutions: a light-weighted NeRF is sufficient, under the guidance of the feature grid representation, to render photorealistic novel views with fine details; and the jointly optimized ground feature planes, can meanwhile gain further refinements, forming a more accurate and compact feature space and output much more natural rendering results.
****
Defining and Quantifying the Emergence of Sparse Concepts in DNNs
Jie Ren, Mingjie Li, Qirui Chen, Huiqi Deng, Quanshi Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20280-20289
This paper aims to illustrate the concept-emerging phenomenon in a trained DNN. Specifically, we find that the inference score of a DNN can be disentangled into the effects of a few interactive concepts. These concepts can be understood as inference patterns in a sparse, symbolic graphical model, which explains the DNN. The faithfulness of using such a graphical model to explain the DNN is theoretically guaranteed, because we prove that the graphical model can well mimic the DNN's outputs on an exponential number of different masked samples. Besides, such a graphical model can be further simplified and re-written as an And-Or graph (AOG), without losing much explanation accuracy. The code is released at https://github.com/sjtu-xai-lab/aog.
****
Uncurated Image-Text Datasets: Shedding Light on Demographic Bias
Noa Garcia, Yusuke Hirota, Yankun Wu, Yuta Nakashima; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6957-6

The increasing tendency to collect large and uncurated datasets to train vision-and-language models has raised concerns about fair representations. It is known that even small but manually annotated datasets, such as MSCOCO, are affected by societal bias. This problem, far from being solved, may be getting worse with data crawled from the Internet without much control. In addition, the lack of tools to analyze societal bias in big collections of images makes addressing the problem extremely challenging. Our first contribution is to annotate part of the Google Conceptual Captions dataset, widely used for training vision-and-language models, with four demographic and two contextual attributes. Our second contribution is to conduct a comprehensive analysis of the annotations, focusing on how different demographic groups are represented. Our last contribution lies in evaluating three prevailing vision-and-language tasks: image captioning, text-image CLIP embeddings, and text-to-image generation, showing that societal bias is a persistent problem in all of them.
****
FreeSeg: Unified, Universal and Open-Vocabulary Image Segmentation
Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, Xingang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19446-19455
Recently, open-vocabulary learning has emerged to accomplish segmentation for arbitrary categories of text-based descriptions, which popularizes the segmentation system to more general-purpose application scenarios. However, existing methods devote to designing specialized architectures or parameters for specific segmentation tasks. These customized design paradigms lead to fragmentation between various segmentation tasks, thus hindering the uniformity of segmentation models. Hence in this paper, we propose FreeSeg, a generic framework to accomplish Unified, Universal and Open-Vocabulary Image Segmentation. FreeSeg optimizes an all-in-one network via one-shot training and employs the same architecture and parameters to handle diverse segmentation tasks seamlessly in the inference procedure. Additionally, adaptive prompt learning facilitates the unified model to capture task-aware and category-sensitive concepts, improving model robustness in multi-task and varied scenarios. Extensive experimental results demonstrate that FreeSeg establishes new state-of-the-art results in performance and generalization on three segmentation tasks, which outperforms the best task-specific architectures by a large margin: 5.5% mIoU on semantic segmentation, 17.6% mAP on instance segmentation, 20.1% PQ on panoptic segmentation for the unseen class on COCO. Project page: https://FreeSeg.github.io.
****
AVFormer: Injecting Vision Into Frozen Speech Models for Zero-Shot AV-ASR
Paul Hongsuck Seo, Arsha Nagrani, Cordelia Schmid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22922-22931
Audiovisual automatic speech recognition (AV-ASR) aims to improve the robustness of a speech recognition system by incorporating visual information. Training fully supervised multimodal models for this task from scratch, however is limited by the need for large labelled audiovisual datasets (in each downstream domain of interest). We present AVFormer, a simple method for augmenting audioonly models with visual information, at the same time performing lightweight domain adaptation. We do this by (i) injecting visual embeddings into a frozen ASR model using lightweight trainable adaptors. We show that these can be trained on a small amount of weakly labelled video data with minimum additional training time and parameters. (ii) We also introduce a simple curriculum scheme during training which we show is crucial to enable the model to jointly process audio and visual information effectively; and finally (iii) we show that our model achieves state of the art zero-shot results on three different AV-ASR benchmarks (How2, VisSpeech and Ego4D), while also crucially preserving decent performance on traditional audio-only speech recognition benchmarks (LibriSpeech). Qualitative results show that our model effectively leverages visual information for robust speech recognition.

****
FreeNeRF: Improving Few-Shot Neural Rendering With Free Frequency Regularization
Jiawei Yang, Marco Pavone, Yue Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8254-8263
Novel view synthesis with sparse inputs is a challenging problem for neural radiance fields (NeRF). Recent efforts alleviate this challenge by introducing external supervision, such as pre-trained models and extra depth signals, or by using non-trivial patch-based rendering. In this paper, we present Frequency regularized NeRF (FreeNeRF), a surprisingly simple baseline that outperforms previous methods with minimal modifications to plain NeRF. We analyze the key challenges in few-shot neural rendering and find that frequency plays an important role in NeRF's training. Based on this analysis, we propose two regularization terms: one to regularize the frequency range of NeRF's inputs, and the other to penalize the near-camera density fields. Both techniques are "free lunches" that come at no additional computational cost. We demonstrate that even with just one line of code change, the original NeRF can achieve similar performance to other complicated methods in the few-shot setting. FreeNeRF achieves state-of-the-art performance across diverse datasets, including Blender, DTU, and LLFF. We hope that this simple baseline will motivate a rethinking of the fundamental role of frequency in NeRF's training, under both the low-data regime and beyond. This project is released at https://jiawei-yang.github.io/FreeNeRF/.
****
Adversarial Robustness via Random Projection Filters
Minjing Dong, Chang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4077-4086
Deep Neural Networks show superior performance in various tasks but are vulnerable to adversarial attacks. Most defense techniques are devoted to the adversarial training strategies, however, it is difficult to achieve satisfactory robust performance only with traditional adversarial training. We mainly attribute it to that aggressive perturbations which lead to the loss increment can always be found via gradient ascent in white-box setting. Although some noises can be involved to prevent attacks from deriving precise gradients on inputs, there exist trade-offs between the defense capability and natural generalization. Taking advantage of the properties of random projection, we propose to replace part of convolutional filters with random projection filters, and theoretically explore the geometric representation preservation of proposed synthesized filters via Johnson-Lindenstrauss lemma. We conduct sufficient evaluation on multiple networks and datasets. The experimental results showcase the superiority of proposed random projection filters to state-of-the-art baselines. The code is available on https://github.com/UniSerj/Random-Projection-Filters.
****
VNE: An Effective Method for Improving Deep Representation by Manipulating Eigenvalue Distribution
Jaeill Kim, Suhyun Kang, Duhun Hwang, Jungwook Shin, Wonjong Rhee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3799-3810
Since the introduction of deep learning, a wide scope of representation properties, such as decorrelation, whitening, disentanglement, rank, isotropy, and mutual information, have been studied to improve the quality of representation. However, manipulating such properties can be challenging in terms of implementational effectiveness and general applicability. To address these limitations, we propose to regularize von Neumann entropy (VNE) of representation. First, we demonstrate that the mathematical formulation of VNE is superior in effectively manipulating the eigenvalues of the representation autocorrelation matrix. Then, we demonstrate that it is widely applicable in improving state-of-the-art algorithms or popular benchmark algorithms by investigating domain-generalization, meta-learning, self-supervised learning, and generative models. In addition, we formally establish theoretical connections with rank, disentanglement, and isotropy of representation. Finally, we provide discussions on the dimension control of VNE and the relationship with Shannon entropy. Code is available at: https://github.com

/jaeill/CVPR23-VNE.
****
Self-Guided Diffusion Models
Vincent Tao Hu, David W. Zhang, Yuki M. Asano, Gertjan J. Burghouts, Cees G. M. Snoek; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18413-18422
Diffusion models have demonstrated remarkable progress in image generation quality, especially when guidance is used to control the generative process. However, guidance requires a large amount of image-annotation pairs for training and is thus dependent on their availability and correctness. In this paper, we eliminate the need for such annotation by instead exploiting the flexibility of self-supervision signals to design a framework for self-guided diffusion models. By leveraging a feature extraction function and a self-annotation function, our method provides guidance signals at various image granularities: from the level of holistic images to object boxes and even segmentation masks. Our experiments on single-label and multi-label image datasets demonstrate that self-labeled guidance always outperforms diffusion models without guidance and may even surpass guidance based on ground-truth labels. When equipped with self-supervised box or mask proposals, our method further generates visually diverse yet semantically consistent images, without the need for any class, box, or segment label annotation. Self-guided diffusion is simple, flexible and expected to profit from deployment at scale.
****
NeuWigs: A Neural Dynamic Model for Volumetric Hair Capture and Animation
Ziyan Wang, Giljoo Nam, Tuur Stuyck, Stephen Lombardi, Chen Cao, Jason Saragih, Michael Zollhöfer, Jessica Hodgins, Christoph Lassner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8641-8651
The capture and animation of human hair are two of the major challenges in the creation of realistic avatars for the virtual reality. Both problems are highly challenging, because hair has complex geometry and appearance, as well as exhibits challenging motion. In this paper, we present a two-stage approach that models hair independently from the head to address these challenges in a data-driven manner. The first stage, state compression, learns a low-dimensional latent space of 3D hair states containing motion and appearance, via a novel autoencoder-as-a-tracker strategy. To better disentangle the hair and head in appearance learning, we employ multi-view hair segmentation masks in combination with a differentiable volumetric renderer. The second stage learns a novel hair dynamics model that performs temporal hair transfer based on the discovered latent codes. To enforce higher stability while driving our dynamics model, we employ the 3D point-cloud autoencoder from the compression stage for de-noising of the hair state. Our model outperforms the state of the art in novel view synthesis and is capable of creating novel hair animations without having to rely on hair observations as a driving signal
****
CLIP2: Contrastive Language-Image-Point Pretraining From Real-World Point Cloud Data
Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, Hang Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15244-15253
Contrastive Language-Image Pre-training, benefiting from large-scale unlabeled text-image pairs, has demonstrated great performance in open-world vision understanding tasks. However, due to the limited Text-3D data pairs, adapting the success of 2D Vision-Language Models (VLM) to the 3D space remains an open problem. Existing works that leverage VLM for 3D understanding generally resort to constructing intermediate 2D representations for the 3D data, but at the cost of losing 3D geometry information. To take a step toward open-world 3D vision understanding, we propose Contrastive Language-Image-Point Cloud Pretraining (CLIP^2) to directly learn the transferable 3D point cloud representation in realistic scenari

os with a novel proxy alignment mechanism. Specifically, we exploit naturally-existed correspondences in 2D and 3D scenarios, and build well-aligned and instance-based text-image-point proxies from those complex scenarios. On top of that, we propose a cross-modal contrastive objective to learn semantic and instance-level aligned point cloud representation. Experimental results on both indoor and outdoor scenarios show that our learned 3D representation has great transfer ability in downstream tasks, including zero-shot and few-shot 3D recognition, which boosts the state-of-the-art methods by large margins. Furthermore, we provide analyses of the capability of different representations in real scenarios and present the optional ensemble scheme.
****
HNeRV: A Hybrid Neural Representation for Videos
Hao Chen, Matthew Gwilliam, Ser-Nam Lim, Abhinav Shrivastava; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10270-10279
Implicit neural representations store videos as neural networks and have performed well for vision tasks such as video compression and denoising. With frame index and/or positional index as input, implicit representations (NeRV, E-NeRV, etc.) reconstruct video frames from fixed and content-agnostic embeddings. Such embedding largely limits the regression capacity and internal generalization for video interpolation. In this paper, we propose a Hybrid Neural Representation for Videos (HNeRV), where learnable and content-adaptive embeddings act as decoder input. Besides the input embedding, we introduce a HNeRV block to make model parameters evenly distributed across the entire network, therefore higher layers (layers near the output) can have more capacity to store high-resolution content and video details. With content-adaptive embedding and re-designed model architecture, HNeRV outperforms implicit methods (NeRV, E-NeRV) in video regression task for both reconstruction quality and convergence speed, and shows better internal generalization. As a simple and efficient video representation, HNeRV also shows decoding advantages for speed, flexibility, and deployment, compared to traditional codecs (H.264, H.265) and learning-based compression methods. Finally, we explore the effectiveness of HNeRV on downstream tasks such as video compression and video inpainting.
****
Model-Agnostic Gender Debiased Image Captioning
Yusuke Hirota, Yuta Nakashima, Noa Garcia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15191-15200
Image captioning models are known to perpetuate and amplify harmful societal bias in the training set. In this work, we aim to mitigate such gender bias in image captioning models. While prior work has addressed this problem by forcing models to focus on people to reduce gender misclassification, it conversely generates gender-stereotypical words at the expense of predicting the correct gender. From this observation, we hypothesize that there are two types of gender bias affecting image captioning models: 1) bias that exploits context to predict gender, and 2) bias in the probability of generating certain (often stereotypical) words because of gender. To mitigate both types of gender biases, we propose a framework, called LIBRA, that learns from synthetically biased samples to decrease both types of biases, correcting gender misclassification and changing gender-stereotypical words to more neutral ones.
****
Local Implicit Ray Function for Generalizable Radiance Field Representation
Xin Huang, Qi Zhang, Ying Feng, Xiaoyu Li, Xuan Wang, Qing Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 97-107
We propose LIRF (Local Implicit Ray Function), a generalizable neural rendering approach for novel view rendering. Current generalizable neural radiance fields (NeRF) methods sample a scene with a single ray per pixel and may therefore render blurred or aliased views when the input views and rendered views observe scene content at different resolutions. To solve this problem, we propose LIRF to aggregate the information from conical frustums to construct a ray. Given 3D posit

ions within conical frustums, LIRF takes 3D coordinates and the features of coni cal frustums as inputs and predicts a local volumetric radiance field. Since the coordinates are continuous, LIRF renders high-quality novel views at a continuo usly-valued scale via volume rendering. Besides, we predict the visible weights for each input view via transformer-based feature matching to improve the perfor mance in occluded areas. Experimental results on real-world scenes validate that our method outperforms state-of-the-art methods on novel view rendering of unse en scenes at arbitrary scales.
****
One-Shot High-Fidelity Talking-Head Synthesis With Deformable Neural Radiance Fi eld
Weichuang Li, Longhao Zhang, Dong Wang, Bin Zhao, Zhigang Wang, Mulin Chen, Bang Zhang, Zhongjian Wang, Liefeng Bo, Xuelong Li; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17969-17978
Talking head generation aims to generate faces that maintain the identity inform ation of the source image and imitate the motion of the driving image. Most pion eering methods rely primarily on 2D representations and thus will inevitably suf fer from face distortion when large head rotations are encountered. Recent works instead employ explicit 3D structural representations or implicit neural render ing to improve performance under large pose changes. Nevertheless, the fidelity of identity and expression is not so desirable, especially for novel-view synthe sis. In this paper, we propose HiDe-NeRF, which achieves high-fidelity and free-view talking-head synthesis. Drawing on the recently proposed Deformable Neural Radiance Fields, HiDe-NeRF represents the 3D dynamic scene into a canonical appe arance field and an implicit deformation field, where the former comprises the c anonical source face and the latter models the driving pose and expression. In p articular, we improve fidelity from two aspects: (i) to enhance identity express iveness, we design a generalized appearance module that leverages multi-scale vo lume features to preserve face shape and details; (ii) to improve expression pre ciseness, we propose a lightweight deformation module that explicitly decouples the pose and expression to enable precise expression modeling. Extensive experim ents demonstrate that our proposed approach can generate better results than pre vious works. Project page: https://www.waytron.net/hidenerf/
****
FitMe: Deep Photorealistic 3D Morphable Model Avatars
Alexandros Lattas, Stylianos Moschoglou, Stylianos Ploumpis, Baris Gecer, Jianka ng Deng, Stefanos Zafeiriou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8629-8640
In this paper, we introduce FitMe, a facial reflectance model and a differentiab le rendering optimization pipeline, that can be used to acquire high-fidelity re nderable human avatars from single or multiple images. The model consists of a m ulti-modal style-based generator, that captures facial appearance in terms of di ffuse and specular reflectance, and a PCA-based shape model. We employ a fast di fferentiable rendering process that can be used in an optimization pipeline, whi le also achieving photorealistic facial shading. Our optimization process accura tely captures both the facial reflectance and shape in high-detail, by exploitin g the expressivity of the style-based latent representation and of our shape mod el. FitMe achieves state-of-the-art reflectance acquisition and identity preserv ation on single "in-the-wild" facial images, while it produces impressive scan-l ike results, when given multiple unconstrained facial images pertaining to the s ame identity. In contrast with recent implicit avatar reconstructions, FitMe req uires only one minute and produces relightable mesh and texture-based avatars, t hat can be used by end-user applications.
****
Dense Distinct Query for End-to-End Object Detection
Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zh ang, Ping Luo, Kai Chen; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2023, pp. 7329-7338
One-to-one label assignment in object detection has successfully obviated the ne ed of non-maximum suppression (NMS) as a postprocessing and makes the pipeline e

nd-to-end. However, it triggers a new dilemma as the widely used sparse queries cannot guarantee a high recall, while dense queries inevitably bring more similar queries and encounters optimization difficulty. As both sparse and dense queries are problematic, then what are the expected queries in end-to-end object detection? This paper shows that the solution should be Dense Distinct Queries (DDQ). Concretely, we first lay dense queries like traditional detectors and then select distinct ones for one-to-one assignments. DDQ blends the advantages of traditional and recent end-to-end detectors and significantly improves the performance of various detectors including FCN, R-CNN, and DETRs. Most impressively, DDQ-DETR achieves 52.1 AP on MS-COCO dataset within 12 epochs using a ResNet-50 backbone, outperforming all existing detectors in the same setting. DDQ also shares the benefit of end-to-end detectors in crowded scenes and achieves 93.8 AP on CrowdHuman. We hope DDQ can inspire researchers to consider the complementarity between traditional methods and end-to-end detectors. The source code can be found at https://github.com/jshilong/DDQ.

****

CLIPPO: Image-and-Language Understanding From Pixels Only

Michael Tschannen, Basil Mustafa, Neil Houlsby; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11006-11017

Multimodal models are becoming increasingly effective, in part due to unified components, such as the Transformer architecture. However, multimodal models still often consist of many task- and modality-specific pieces and training procedures. For example, CLIP (Radford et al., 2021) trains independent text and image towers via a contrastive loss. We explore an additional unification: the use of a pure pixel-based model to perform image, text, and multimodal tasks. Our model is trained with contrastive loss alone, so we call it CLIP-Pixels Only (CLIPPO). CLIPPO uses a single encoder that processes both regular images and text rendered as images. CLIPPO performs image-based tasks such as retrieval and zero-shot image classification almost as well as CLIP-style models, with half the number of parameters and no text-specific tower or embedding. When trained jointly via image-text contrastive learning and next-sentence contrastive learning, CLIPPO can perform well on natural language understanding tasks, without any word-level loss (language modelling or masked language modelling), outperforming pixel-based prior work. Surprisingly, CLIPPO can obtain good accuracy in visual question answering, simply by rendering the question and image together. Finally, we exploit the fact that CLIPPO does not require a tokenizer to show that it can achieve strong performance on multilingual multimodal retrieval without modifications. Code and pretrained models are available at https://github.com/google-research/big_vision.

****

Trajectory-Aware Body Interaction Transformer for Multi-Person Pose Forecasting

Xiaogang Peng, Siyuan Mao, Zizhao Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17121-17130

Multi-person pose forecasting remains a challenging problem, especially in modeling fine-grained human body interaction in complex crowd scenarios. Existing methods typically represent the whole pose sequence as a temporal series, yet overlook interactive influences among people based on skeletal body parts. In this paper, we propose a novel Trajectory-Aware Body Interaction Transformer (TBIFormer) for multi-person pose forecasting via effectively modeling body part interactions. Specifically, we construct a Temporal Body Partition Module that transforms all the pose sequences into a Multi-Person Body-Part sequence to retain spatial and temporal information based on body semantics. Then, we devise a Social Body Interaction Self-Attention (SBI-MSA) module, utilizing the transformed sequence to learn body part dynamics for inter- and intra-individual interactions. Furthermore, different from prior Euclidean distance-based spatial encodings, we present a novel and efficient Trajectory-Aware Relative Position Encoding for SBI-MSA to offer discriminative spatial information and additional interactive clues. On both short- and long-term horizons, we empirically evaluate our framework on CMU-Mocap, MuPoTS-3D as well as synthesized datasets (6   10 persons), and demonstrate that our method greatly outperforms the state-of-the-art methods.

****
Conditional Image-to-Video Generation With Latent Flow Diffusion Models
Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, Martin Renqiang Min; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18444-18455
Conditional image-to-video (cI2V) generation aims to synthesize a new plausible video starting from an image (e.g., a person's face) and a condition (e.g., an action class label like smile). The key challenge of the cI2V task lies in the simultaneous generation of realistic spatial appearance and temporal dynamics corresponding to the given image and condition. In this paper, we propose an approach for cI2V using novel latent flow diffusion models (LFDM) that synthesize an optical flow sequence in the latent space based on the given condition to warp the given image. Compared to previous direct-synthesis-based works, our proposed LFDM can better synthesize spatial details and temporal motion by fully utilizing the spatial content of the given image and warping it in the latent space according to the generated temporally-coherent flow. The training of LFDM consists of two separate stages: (1) an unsupervised learning stage to train a latent flow auto-encoder for spatial content generation, including a flow predictor to estimate latent flow between pairs of video frames, and (2) a conditional learning stage to train a 3D-UNet-based diffusion model (DM) for temporal latent flow generation. Unlike previous DMs operating in pixel space or latent feature space that couples spatial and temporal information, the DM in our LFDM only needs to learn a low-dimensional latent flow space for motion generation, thus being more computationally efficient. We conduct comprehensive experiments on multiple datasets, where LFDM consistently outperforms prior arts. Furthermore, we show that LFDM can be easily adapted to new domains by simply finetuning the image decoder. Our code is available at https://github.com/nihaomiao/CVPR23_LFDM.
****
Virtual Sparse Convolution for Multimodal 3D Object Detection
Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21653-21662
Recently, virtual/pseudo-point-based 3D object detection that seamlessly fuses RGB images and LiDAR data by depth completion has gained great attention. However, virtual points generated from an image are very dense, introducing a huge amount of redundant computation during detection. Meanwhile, noises brought by inaccurate depth completion significantly degrade detection precision. This paper proposes a fast yet effective backbone, termed VirConvNet, based on a new operator VirConv (Virtual Sparse Convolution), for virtual-point-based 3D object detection. The VirConv consists of two key designs: (1) StVD (Stochastic Voxel Discard) and (2) NRConv (Noise-Resistant Submanifold Convolution). The StVD alleviates the computation problem by discarding large amounts of nearby redundant voxels. The NRConv tackles the noise problem by encoding voxel features in both 2D image and 3D LiDAR space. By integrating our VirConv, we first develop an efficient pipeline VirConv-L based on an early fusion design. Then, we build a high-precision pipeline VirConv-T based on a transformed refinement scheme. Finally, we develop a semi-supervised pipeline VirConv-S based on a pseudo-label framework. On the KITTI car 3D detection test leaderboard, our VirConv-L achieves 85% AP with a fast running speed of 56ms. Our VirConv-T and VirConv-S attains a high-precision of 86.3% and 87.2% AP, and currently rank 2nd and 1st, respectively. The code is available at https://github.com/hailanyi/VirConv.
****
DETR With Additional Global Aggregation for Cross-Domain Weakly Supervised Object Detection
Zongheng Tang, Yifan Sun, Si Liu, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11422-11432
This paper presents a DETR-based method for cross-domain weakly supervised object detection (CDWSOD), aiming at adapting the detector from source to target domain through weak supervision. We think DETR has strong potential for CDWSOD due to an insight: the encoder and the decoder in DETR are both based on the attentio

n mechanism and are thus capable of aggregating semantics across the entire image. The aggregation results, i.e., image-level predictions, can naturally exploit the weak supervision for domain alignment. Such motivated, we propose DETR with additional Global Aggregation (DETR-GA), a CDWSOD detector that simultaneously makes "instance-level + image-level" predictions and utilizes "strong + weak" supervisions. The key point of DETR-GA is very simple: for the encoder / decoder, we respectively add multiple class queries / a foreground query to aggregate the semantics into image-level predictions. Our query-based aggregation has two advantages. First, in the encoder, the weakly-supervised class queries are capable of roughly locating the corresponding positions and excluding the distraction from non-relevant regions. Second, through our design, the object queries and the foreground query in the decoder share consensus on the class semantics, therefore making the strong and weak supervision mutually benefit each other for domain alignment. Extensive experiments on four popular cross-domain benchmarks show that DETR-GA significantly improves CSWSOD and advances the states of the art (e.g., 29.0% --> 79.4% mAP on PASCAL VOC --> Clipart_all dataset).
****
Divide and Adapt: Active Domain Adaptation via Customized Learning
Duojun Huang, Jichang Li, Weikai Chen, Junshi Huang, Zhenhua Chai, Guanbin Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7651-7660
Active domain adaptation (ADA) aims to improve the model adaptation performance by incorporating the active learning (AL) techniques to label a maximally-informative subset of target samples. Conventional AL methods do not consider the existence of domain shift, and hence, fail to identify the truly valuable samples in the context of domain adaptation. To accommodate active learning and domain adaption, the two naturally different tasks, in a collaborative framework, we advocate that a customized learning strategy for the target data is the key to the success of ADA solutions. We present Divide-and-Adapt (DiaNA), a new ADA framework that partitions the target instances into four categories with stratified transferable properties. With a novel data subdivision protocol based on uncertainty and domainness, DiaNA can accurately recognize the most gainful samples. While sending the informative instances for annotation, DiaNA employs tailored learning strategies for the remaining categories. Furthermore, we propose an informativeness score that unifies the data partitioning criteria. This enables the use of a Gaussian mixture model (GMM) to automatically sample unlabeled data into the proposed four categories. Thanks to the "divide-and-adapt" spirit, DiaNA can handle data with large variations of domain gap. In addition, we show that DiaNA can generalize to different domain adaptation settings, such as unsupervised domain adaptation (UDA), semi-supervised domain adaptation (SSDA), source-free domain adaptation (SFDA), etc.
****
Towards Universal Fake Image Detectors That Generalize Across Generative Models
Utkarsh Ojha, Yuheng Li, Yong Jae Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24480-24489
With generative models proliferating at a rapid rate, there is a growing need for general purpose fake image detectors. In this work, we first show that the existing paradigm, which consists of training a deep network for real-vs-fake classification, fails to detect fake images from newer breeds of generative models when trained to detect GAN fake images. Upon analysis, we find that the resulting classifier is asymmetrically tuned to detect patterns that make an image fake. The real class becomes a 'sink' class holding anything that is not fake, including generated images from models not accessible during training. Building upon this discovery, we propose to perform real-vs-fake classification without learning; i.e., using a feature space not explicitly trained to distinguish real from fake images. We use nearest neighbor and linear probing as instantiations of this idea. When given access to the feature space of a large pretrained vision-language model, the very simple baseline of nearest neighbor classification has surprisingly good generalization ability in detecting fake images from a wide variety of generative models; e.g., it improves upon the SoTA by +15.07 mAP and +25.90% a

cc when tested on unseen diffusion and autoregressive models.
****
Towards Bridging the Performance Gaps of Joint Energy-Based Models
Xiulong Yang, Qing Su, Shihao Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15732-15741
Can we train a hybrid discriminative-generative model with a single network? This question has recently been answered in the affirmative, introducing the field of Joint Energy-based Model (JEM), which achieves high classification accuracy and image generation quality simultaneously. Despite recent advances, there remain two performance gaps: the accuracy gap to the standard softmax classifier, and the generation quality gap to state-of-the-art generative models. In this paper, we introduce a variety of training techniques to bridge the accuracy gap and the generation quality gap of JEM. 1) We incorporate a recently proposed sharpness-aware minimization (SAM) framework to train JEM, which promotes the energy landscape smoothness and the generalization of JEM. 2) We exclude data augmentation from the maximum likelihood estimate pipeline of JEM, and mitigate the negative impact of data augmentation to image generation quality. Extensive experiments on multiple datasets demonstrate our SADA-JEM achieves state-of-the-art performances and outperforms JEM in image classification, image generation, calibration, out-of-distribution detection and adversarial robustness by a notable margin. Our code is available at https://github.com/sndnyang/SADAJEM.
****
Learning Spatial-Temporal Implicit Neural Representations for Event-Guided Video Super-Resolution
Yunfan Lu, Zipeng Wang, Minjie Liu, Hongjian Wang, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1557-1567
Event cameras sense the intensity changes asynchronously and produce event streams with high dynamic range and low latency. This has inspired research endeavors utilizing events to guide the challenging video super-resolution (VSR) task. In this paper, we make the first at tempt to address a novel problem of achieving VSR at random scales by taking advantages of the high temporal resolution property of events. This is hampered by the difficulties of representing the spatial-temporal information of events when guiding VSR. To this end, we propose a novel framework that incorporates the spatial-temporal interpolation of events to VSR in a unified framework. Our key idea is to learn implicit neural representations from queried spatial-temporal coordinates and features from both RGB frames and events. Our method contains three parts. Specifically, the Spatial-Temporal Fusion (STF) module first learns the 3D features from events and RGB frames. Then, the Temporal Filter (TF) module unlocks more explicit motion information from the events near the queried timestamp and generates the 2D features. Lastly, the Spatial-Temporal Implicit Representation (STIR) module recovers the SR frame in arbitrary resolutions from the outputs of these two modules. In addition, we collect a real-world dataset with spatially aligned events and RGB frames. Extensive experiments show that our method significantly surpass the prior-arts and achieves VSR with random scales, e.g., 6.5. Code and dataset are available at https://.
****
Both Style and Distortion Matter: Dual-Path Unsupervised Domain Adaptation for Panoramic Semantic Segmentation
Xu Zheng, Jinjing Zhu, Yexin Liu, Zidong Cao, Chong Fu, Lin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1285-1295
The ability of scene understanding has sparked active research for panoramic image semantic segmentation. However, the performance is hampered by distortion of the equirectangular projection (ERP) and a lack of pixel-wise annotations. For this reason, some works treat the ERP and pinhole images equally and transfer knowledge from the pinhole to ERP images via unsupervised domain adaptation (UDA). However, they fail to handle the domain gaps caused by: 1) the inherent differences between camera sensors and captured scenes; 2) the distinct image formats (e

.g., ERP and pinhole images). In this paper, we propose a novel yet flexible dual-path UDA framework, DPPASS, taking ERP and tangent projection (TP) images as inputs. To reduce the domain gaps, we propose cross-projection and intra-projection training. The cross-projection training includes tangent-wise feature contrastive training and prediction consistency training. That is, the former formulates the features with the same projection locations as positive examples and vice versa, for the models' awareness of distortion, while the latter ensures the consistency of cross-model predictions between the ERP and TP. Moreover, adversarial intra-projection training is proposed to reduce the inherent gap, between the features of the pinhole images and those of the ERP and TP images, respectively. Importantly, the TP path can be freely removed after training, leading to no additional inference cost. Extensive experiments on two benchmarks show that our DPPASS achieves +1.06% mIoU increment than the state-of-the-art approaches.
****
expOSE: Accurate Initialization-Free Projective Factorization Using Exponential Regularization
José Pedro Iglesias, Amanda Nilsson, Carl Olsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8959-8968
Bundle adjustment is a key component in practically all available Structure from Motion systems. While it is crucial for achieving accurate reconstruction, convergence to the right solution hinges on good initialization. The recently introduced factorization-based pOSE methods formulate a surrogate for the bundle adjustment error without reliance on good initialization. In this paper, we show that pOSE has an undesirable penalization of large depths. To address this we propose expOSE which has an exponential regularization that is negligible for positive depths. To achieve efficient inference we use a quadratic approximation that allows an iterative solution with VarPro. Furthermore, we extend the method with radial distortion robustness by decomposing the Object Space Error into radial and tangential components. Experimental results confirm that the proposed method is robust to initialization and improves reconstruction quality compared to state-of-the-art methods even without bundle adjustment refinement.
****
OpenGait: Revisiting Gait Recognition Towards Better Practicality
Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, Shiqi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9707-9716
Gait recognition is one of the most critical long-distance identification technologies and increasingly gains popularity in both research and industry communities. Despite the significant progress made in indoor datasets, much evidence shows that gait recognition techniques perform poorly in the wild. More importantly, we also find that some conclusions drawn from indoor datasets cannot be generalized to real applications. Therefore, the primary goal of this paper is to present a comprehensive benchmark study for better practicality rather than only a particular model for better performance. To this end, we first develop a flexible and efficient gait recognition codebase named OpenGait. Based on OpenGait, we deeply revisit the recent development of gait recognition by re-conducting the ablative experiments. Encouragingly,we detect some unperfect parts of certain prior woks, as well as new insights. Inspired by these discoveries, we develop a structurally simple, empirically powerful, and practically robust baseline model, GaitBase. Experimentally, we comprehensively compare GaitBase with many current gait recognition methods on multiple public datasets, and the results reflect that GaitBase achieves significantly strong performance in most cases regardless of indoor or outdoor situations. Code is available at https://github.com/ShiqiYu/OpenGait.
****
ALTO: Alternating Latent Topologies for Implicit 3D Reconstruction
Zhen Wang, Shijie Zhou, Jeong Joon Park, Despoina Paschalidou, Suya You, Gordon Wetzstein, Leonidas Guibas, Achuta Kadambi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 259-270
This work introduces alternating latent topologies (ALTO) for high-fidelity reco

nstruction of implicit 3D surfaces from noisy point clouds. Previous work identifies that the spatial arrangement of latent encodings is important to recover detail. One school of thought is to encode a latent vector for each point (point latents). Another school of thought is to project point latents into a grid (grid latents) which could be a voxel grid or triplane grid. Each school of thought has tradeoffs. Grid latents are coarse and lose high-frequency detail. In contrast, point latents preserve detail. However, point latents are more difficult to decode into a surface, and quality and runtime suffer. In this paper, we propose ALTO to sequentially alternate between geometric representations, before converging to an easy-to-decode latent. We find that this preserves spatial expressiveness and makes decoding lightweight. We validate ALTO on implicit 3D recovery and observe not only a performance improvement over the state-of-the-art, but a runtime improvement of 3-10x. Anonymized source code at https://visual.ee.ucla.edu/alto.htm/.
****

Learning Debiased Representations via Conditional Attribute Interpolation
Yi-Kai Zhang, Qi-Wei Wang, De-Chuan Zhan, Han-Jia Ye; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7599-7608
An image is usually described by more than one attribute like "shape" and "color". When a dataset is biased, i.e., most samples have attributes spuriously correlated with the target label, a Deep Neural Network (DNN) is prone to make predictions by the "unintended" attribute, especially if it is easier to learn. To improve the generalization ability when training on such a biased dataset, we propose a chi^2-model to learn debiased representations. First, we design a chi-shape pattern to match the training dynamics of a DNN and find Intermediate Attribute Samples (IASs) --- samples near the attribute decision boundaries, which indicate how the value of an attribute changes from one extreme to another. Then we rectify the representation with a chi-structured metric learning objective. Conditional interpolation among IASs eliminates the negative effect of peripheral attributes and facilitates retaining the intra-class compactness. Experiments show that chi^2-model learns debiased representation effectively and achieves remarkable improvements on various datasets.
****

A Large-Scale Homography Benchmark
Daniel Barath, Dmytro Mishkin, Michal Polic, Wolfgang Förstner, Jiri Matas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21360-21370
We present a large-scale dataset of Planes in 3D, Pi3D, of roughly 1000 planes observed in 10 000 images from the 1DSfM dataset, and HEB, a large-scale homography estimation benchmark leveraging Pi3D. The applications of the Pi3D dataset are diverse, e.g. training or evaluating monocular depth, surface normal estimation and image matching algorithms. The HEB dataset consists of 226 260 homographies and includes roughly 4M correspondences. The homographies link images that often undergo significant viewpoint and illumination changes. As applications of HEB, we perform a rigorous evaluation of a wide range of robust estimators and deep learning-based correspondence filtering methods, establishing the current state-of-the-art in robust homography estimation. We also evaluate the uncertainty of the SIFT orientations and scales w.r.t. the ground truth coming from the underlying homographies and provide codes for comparing uncertainty of custom detectors.
****

Modeling Inter-Class and Intra-Class Constraints in Novel Class Discovery
Wenbin Li, Zhichen Fan, Jing Huo, Yang Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3449-3458
Novel class discovery (NCD) aims at learning a model that transfers the common knowledge from a class-disjoint labelled dataset to another unlabelled dataset and discovers new classes (clusters) within it. Many methods, as well as elaborate training pipelines and appropriate objectives, have been proposed and considerably boosted performance on NCD tasks. Despite all this, we find that the existin

g methods do not sufficiently take advantage of the essence of the NCD setting. To this end, in this paper, we propose to model both inter-class and intra-class constraints in NCD based on the symmetric Kullback-Leibler divergence (sKLD). Specifically, we propose an inter-class sKLD constraint to effectively exploit the disjoint relationship between labelled and unlabelled classes, enforcing the separability for different classes in the embedding space. In addition, we present an intra-class sKLD constraint to explicitly constrain the intra-relationship between a sample and its augmentations and ensure the stability of the training process at the same time. We conduct extensive experiments on the popular CIFAR10, CIFAR100 and ImageNet benchmarks and successfully demonstrate that our method can establish a new state of the art and can achieve significant performance improvements, e.g., 3.5%/3.7% clustering accuracy improvements on CIFAR100-50 dataset split under the task-aware/-agnostic evaluation protocol, over previous state-of-the-art methods. Code is available at https://github.com/FanZhichen/NCD-IIC.
****
Weakly Supervised Video Emotion Detection and Prediction via Cross-Modal Temporal Erasing Network
Zhicheng Zhang, Lijuan Wang, Jufeng Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18888-18897
Automatically predicting the emotions of user-generated videos (UGVs) receives increasing interest recently. However, existing methods mainly focus on a few key visual frames, which may limit their capacity to encode the context that depicts the intended emotions. To tackle that, in this paper, we propose a cross-modal temporal erasing network that locates not only keyframes but also context and audio-related information in a weakly-supervised manner. In specific, we first leverage the intra- and inter-modal relationship among different segments to accurately select keyframes. Then, we iteratively erase keyframes to encourage the model to concentrate on the contexts that include complementary information. Extensive experiments on three challenging video emotion benchmarks demonstrate that our method performs favorably against state-of-the-art approaches. The code is released on https://github.com/nku-zhichengzhang/WECL.
****
Multiple Instance Learning via Iterative Self-Paced Supervised Contrastive Learning
Kangning Liu, Weicheng Zhu, Yiqiu Shen, Sheng Liu, Narges Razavian, Krzysztof J. Geras, Carlos Fernandez-Granda; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3355-3365
Learning representations for individual instances when only bag-level labels are available is a fundamental challenge in multiple instance learning (MIL). Recent works have shown promising results using contrastive self-supervised learning (CSSL), which learns to push apart representations corresponding to two different randomly-selected instances. Unfortunately, in real-world applications such as medical image classification, there is often class imbalance, so randomly-selected instances mostly belong to the same majority class, which precludes CSSL from learning inter-class differences. To address this issue, we propose a novel framework, Iterative Self-paced Supervised Contrastive Learning for MIL Representations (ItS2CLR), which improves the learned representation by exploiting instance-level pseudo labels derived from the bag-level labels. The framework employs a novel self-paced sampling strategy to ensure the accuracy of pseudo labels. We evaluate ItS2CLR on three medical datasets, showing that it improves the quality of instance-level pseudo labels and representations, and outperforms existing MIL methods in terms of both bag and instance level accuracy. Code is available at https://github.com/Kangningthu/ItS2CLR
****
Consistent View Synthesis With Pose-Guided Diffusion Models
Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, Johannes Kopf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16773-16783
Novel view synthesis from a single image has been a cornerstone problem for many

Virtual Reality applications that provide immersive experiences. However, most existing techniques can only synthesize novel views within a limited range of camera motion or fail to generate consistent and high-quality novel views under significant camera movement. In this work, we propose a pose-guided diffusion model to generate a consistent long-term video of novel views from a single image. We design an attention layer that uses epipolar lines as constraints to facilitate the association between different viewpoints. Experimental results on synthetic and real-world datasets demonstrate the effectiveness of the proposed diffusion model against state-of-the-art transformer-based and GAN-based approaches. More qualitative results are available at https://poseguided-diffusion.github.io/.
****
MSMDFusion: Fusing LiDAR and Camera at Multiple Scales With Multi-Depth Seeds for 3D Object Detection
Yang Jiao, Zequn Jie, Shaoxiang Chen, Jingjing Chen, Lin Ma, Yu-Gang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21643-21652
Fusing LiDAR and camera information is essential for accurate and reliable 3D object detection in autonomous driving systems. This is challenging due to the difficulty of combining multi-granularity geometric and semantic features from two drastically different modalities. Recent approaches aim at exploring the semantic densities of camera features through lifting points in 2D camera images (referred to as "seeds") into 3D space, and then incorporate 2D semantics via cross-modal interaction or fusion techniques. However, depth information is under-investigated in these approaches when lifting points into 3D space, thus 2D semantics can not be reliably fused with 3D points. Moreover, their multi-modal fusion strategy, which is implemented as concatenation or attention, either can not effectively fuse 2D and 3D information or is unable to perform fine-grained interactions in the voxel space. To this end, we propose a novel framework with better utilization of the depth information and fine-grained cross-modal interaction between LiDAR and camera, which consists of two important components. First, a Multi-Depth Unprojection (MDU) method is used to enhance the depth quality of the lifted points at each interaction level. Second, a Gated Modality-Aware Convolution (GMA-Conv) block is applied to modulate voxels involved with the camera modality in a fine-grained manner and then aggregate multi-modal features into a unified space. Together they provide the detection head with more comprehensive features from LiDAR and camera. On the nuScenes test benchmark, our proposed method, abbreviated as MSMDFusion, achieves state-of-the-art results on both 3D object detection and tracking tasks without using test-time-augmentation and ensemble techniques. The code is available at https://github.com/SxJyJay/MSMDFusion.
****
Dense-Localizing Audio-Visual Events in Untrimmed Videos: A Large-Scale Benchmark and Baseline
Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, Feng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22942-22951
Existing audio-visual event localization (AVE) handles manually trimmed videos with only a single instance in each of them. However, this setting is unrealistic as natural videos often contain numerous audio-visual events with different categories. To better adapt to real-life applications, in this paper we focus on the task of dense-localizing audio-visual events, which aims to jointly localize and recognize all audio-visual events occurring in an untrimmed video. The problem is challenging as it requires fine-grained audio-visual scene and context understanding. To tackle this problem, we introduce the first Untrimmed Audio-Visual (UnAV-100) dataset, which contains 10K untrimmed videos with over 30K audio-visual events. Each video has 2.8 audio-visual events on average, and the events are usually related to each other and might co-occur as in real-life scenes. Next, we formulate the task using a new learning-based framework, which is capable of fully integrating audio and visual modalities to localize audio-visual events with various lengths and capture dependencies between them in a single pass. Extensive experiments demonstrate the effectiveness of our method as well as the sig

nificance of multi-scale cross-modal perception and dependency modeling for this task.
****

Weak-Shot Object Detection Through Mutual Knowledge Transfer

Xuanyi Du, Weitao Wan, Chong Sun, Chen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19671-19680

Weak-shot Object Detection methods exploit a fully-annotated source dataset to facilitate the detection performance on the target dataset which only contains image-level labels for novel categories. To bridge the gap between these two datasets, we aim to transfer the object knowledge between the source (S) and target (T) datasets in a bi-directional manner. We propose a novel Knowledge Transfer (KT) loss which simultaneously distills the knowledge of objectness and class entropy from a proposal generator trained on the S dataset to optimize a multiple instance learning module on the T dataset. By jointly optimizing the classification loss and the proposed KT loss, the multiple instance learning module effectively learns to classify object proposals into novel categories in the T dataset with the transferred knowledge from base categories in the S dataset. Noticing the predicted boxes on the T dataset can be regarded as an extension for the original annotations on the S dataset to refine the proposal generator in return, we further propose a novel Consistency Filtering (CF) method to reliably remove inaccurate pseudo labels by evaluating the stability of the multiple instance learning module upon noise injections. Via mutually transferring knowledge between the S and T datasets in an iterative manner, the detection performance on the target dataset is significantly improved. Extensive experiments on public benchmarks validate that the proposed method performs favourably against the state-of-the-art methods without increasing the model parameters or inference computational complexity.
****

DATID-3D: Diversity-Preserved Domain Adaptation Using Text-to-Image Diffusion for 3D Generative Model

Gwanghyun Kim, Se Young Chun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14203-14213

Recent 3D generative models have achieved remarkable performance in synthesizing high resolution photorealistic images with view consistency and detailed 3D shapes, but training them for diverse domains is challenging since it requires massive training images and their camera distribution information. Text-guided domain adaptation methods have shown impressive performance on converting the 2D generative model on one domain into the models on other domains with different styles by leveraging the CLIP (Contrastive Language-Image Pre-training), rather than collecting massive datasets for those domains. However, one drawback of them is that the sample diversity in the original generative model is not well-preserved in the domain-adapted generative models due to the deterministic nature of the CLIP text encoder. Text-guided domain adaptation will be even more challenging for 3D generative models not only because of catastrophic diversity loss, but also because of inferior text-image correspondence and poor image quality. Here we propose DATID-3D, a domain adaptation method tailored for 3D generative models using text-to-image diffusion models that can synthesize diverse images per text prompt without collecting additional images and camera information for the target domain. Unlike 3D extensions of prior text-guided domain adaptation methods, our novel pipeline was able to fine-tune the state-of-the-art 3D generator of the source domain to synthesize high resolution, multi-view consistent images in text-guided targeted domains without additional data, outperforming the existing text-guided domain adaptation methods in diversity and text-image correspondence. Furthermore, we propose and demonstrate diverse 3D image manipulations such as one-shot instance-selected adaptation and single-view manipulated 3D reconstruction to fully enjoy diversity in text.
****

CrowdCLIP: Unsupervised Crowd Counting via Vision-Language Model

Dingkang Liang, Jiahao Xie, Zhikang Zou, Xiaoqing Ye, Wei Xu, Xiang Bai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR

), 2023, pp. 2893-2903
Supervised crowd counting relies heavily on costly manual labeling, which is difficult and expensive, especially in dense scenes. To alleviate the problem, we propose a novel unsupervised framework for crowd counting, named CrowdCLIP. The core idea is built on two observations: 1) the recent contrastive pre-trained vision-language model (CLIP) has presented impressive performance on various downstream tasks; 2) there is a natural mapping between crowd patches and count text. To the best of our knowledge, CrowdCLIP is the first to investigate the vision-language knowledge to solve the counting problem. Specifically, in the training stage, we exploit the multi-modal ranking loss by constructing ranking text prompts to match the size-sorted crowd patches to guide the image encoder learning. In the testing stage, to deal with the diversity of image patches, we propose a simple yet effective progressive filtering strategy to first select the highly potential crowd patches and then map them into the language space with various counting intervals. Extensive experiments on five challenging datasets demonstrate that the proposed CrowdCLIP achieves superior performance compared to previous unsupervised state-of-the-art counting methods. Notably, CrowdCLIP even surpasses some popular fully-supervised methods under the cross-dataset setting. The source code will be available at https://github.com/dk-liang/CrowdCLIP.
****
Toward Stable, Interpretable, and Lightweight Hyperspectral Super-Resolution
Wen-jin Guo, Weiying Xie, Kai Jiang, Yunsong Li, Jie Lei, Leyuan Fang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22272-22281
For real applications, existing HSI-SR methods are mostly not only limited to unstable performance under unknown scenarios but also suffer from high computation consumption. In this paper, we develop a new coordination optimization framework for stable, interpretable, and lightweight HSI-SR. Specifically, we create a positive cycle between fusion and degradation estimation under a new probabilistic framework. The estimated degradation is applied to fusion as guidance for a degradation-aware HSI-SR. Under the framework, we establish an explicit degradation estimation method to tackle the indeterminacy and unstable performance driven by black-box simulation in previous methods. Considering the interpretability in fusion, we integrate spectral mixing prior to the fusion process, which can be easily realized by a tiny autoencoder, leading to a dramatic release of the computation burden. We then develop a partial fine-tune strategy in inference to reduce the computation cost further. Comprehensive experiments demonstrate the superiority of our method against state-of-the-art under synthetic and real datasets. For instance, we achieve a 2.3 dB promotion on PSNR with 120x model size reduction and 4300x FLOPs reduction under the CAVE dataset. Code is available in https://github.com/WenjinGuo/DAEM.
****
Masked Auto-Encoders Meet Generative Adversarial Networks and Beyond
Zhengcong Fei, Mingyuan Fan, Li Zhu, Junshi Huang, Xiaoming Wei, Xiaolin Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24449-24459
Masked Auto-Encoder (MAE) pretraining methods randomly mask image patches and then train a vision Transformer to reconstruct the original pixels based on the unmasked patches. While they demonstrates impressive performance for downstream vision tasks, it generally requires a large amount of training resource. In this paper, we introduce a novel Generative Adversarial Networks alike framework, referred to as GAN-MAE, where a generator is used to generate the masked patches according to the remaining visible patches, and a discriminator is employed to predict whether the patch is synthesized by the generator. We believe this capacity of distinguishing whether the image patch is predicted or original is benefit to representation learning. Another key point lies in that the parameters of the vision Transformer backbone in the generator and discriminator are shared. Extensive experiments demonstrate that adversarial training of GAN-MAE framework is more efficient and accordingly outperforms the standard MAE given the same model size, training data, and computation resource. The gains are substantially robust

 for different model sizes and datasets, in particular, a ViT-B model trained wi
th GAN-MAE for 200 epochs outperforms the MAE with 1600 epochs on fine-tuning to
p-1 accuracy of ImageNet-1k with much less FLOPs. Besides, our approach also wor
ks well at transferring downstream tasks.
****
iCLIP: Bridging Image Classification and Contrastive Language-Image Pre-Training
 for Visual Recognition
Yixuan Wei, Yue Cao, Zheng Zhang, Houwen Peng, Zhuliang Yao, Zhenda Xie, Han Hu,
 Baining Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 2776-2786
This paper presents a method that effectively combines two prevalent visual reco
gnition methods, i.e., image classification and contrastive language-image pre-t
raining, dubbed iCLIP. Instead of naive multi-task learning that use two separat
e heads for each task, we fuse the two tasks in a deep fashion that adapts the i
mage classification to share the same formula and the same model weights with th
e language-image pre-training. To further bridge these two tasks, we propose to
enhance the category names in image classification tasks using external knowledg
e, such as their descriptions in dictionaries. Extensive experiments show that t
he proposed method combines the advantages of two tasks well: the strong discrim
ination ability in image classification tasks due to the clear and clean categor
y labels, and the good zero-shot ability in CLIP tasks ascribed to the richer se
mantics in the text descriptions. In particular, it reaches 82.9% top-1 accuracy
 on IN-1K, and surpasses CLIPby 1.8%, with similar model size, on zero-shot reco
gnition of Kornblith 12-dataset benchmark. The code and models are publicly avai
lable at https://github.com/weiyx16/iCLIP.
****
Learning Neural Volumetric Representations of Dynamic Humans in Minutes
Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, Xiaowei Zhou; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8759
-8770
This paper addresses the challenge of efficiently reconstructing volumetric vide
os of dynamic humans from sparse multi-view videos. Some recent works represent
a dynamic human as a canonical neural radiance field (NeRF) and a motion field,
which are learned from input videos through differentiable rendering. But the pe
r-scene optimization generally requires hours. Other generalizable NeRF models l
everage learned prior from datasets to reduce the optimization time by only fine
tuning on new scenes at the cost of visual fidelity. In this paper, we propose a
 novel method for learning neural volumetric representations of dynamic humans i
n minutes with competitive visual quality. Specifically, we define a novel part-
based voxelized human representation to better distribute the representational p
ower of the network to different human parts. Furthermore, we propose a novel 2D
 motion parameterization scheme to increase the convergence rate of deformation
field learning. Experiments demonstrate that our model can be learned 100 times
faster than previous per-scene optimization methods while being competitive in t
he rendering quality. Training our model on a 512x512 video with 100 frames typi
cally takes about 5 minutes on a single RTX 3090 GPU. The code is available on o
ur project page: https://zju3dv.github.io/instant_nvr
****
Streaming Video Model
Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, Zheng-Jun Z
ha; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogn
ition (CVPR), 2023, pp. 14602-14612
Video understanding tasks have traditionally been modeled by two separate archit
ectures, specially tailored for two distinct tasks. Sequence-based video tasks,
such as action recognition, use a video backbone to directly extract spatiotempo
ral features, while frame-based video tasks, such as multiple object tracking (M
OT), rely on single fixed-image backbone to extract spatial features. In contras
t, we propose to unify video understanding tasks into one novel streaming video
architecture, referred to as Streaming Vision Transformer (S-ViT). S-ViT first p
roduces frame-level features with a memory-enabled temporally-aware spatial enco

der to serve the frame-based video tasks. Then the frame features are input into a task-related temporal decoder to obtain spatiotemporal features for sequence-based tasks. The efficiency and efficacy of S-ViT is demonstrated by the state-of-the-art accuracy in the sequence-based action recognition task and the competitive advantage over conventional architecture in the frame-based MOT task. We believe that the concept of streaming video model and the implementation of S-ViT are solid steps towards a unified deep learning architecture for video understanding. Code will be available at https://github.com/yuzhms/Streaming-Video-Model.
****
CapDet: Unifying Dense Captioning and Open-World Detection Pretraining
Yanxin Long, Youpeng Wen, Jianhua Han, Hang Xu, Pengzhen Ren, Wei Zhang, Shen Zhao, Xiaodan Liang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15233-15243
Benefiting from large-scale vision-language pre-training on image-text pairs, open-world detection methods have shown superior generalization ability under the zero-shot or few-shot detection settings. However, a pre-defined category space is still required during the inference stage of existing methods and only the objects belonging to that space will be predicted. To introduce a "real" open-world detector, in this paper, we propose a novel method named CapDet to either predict under a given category list or directly generate the category of predicted bounding boxes. Specifically, we unify the open-world detection and dense caption tasks into a single yet effective framework by introducing an additional dense captioning head to generate the region-grounded captions. Besides, adding the captioning task will in turn benefit the generalization of detection performance since the captioning dataset covers more concepts. Experiment results show that by unifying the dense caption task, our CapDet has obtained significant performance improvements (e.g., +2.1% mAP on LVIS rare classes) over the baseline method on LVIS (1203 classes). Besides, our CapDet also achieves state-of-the-art performance on dense captioning tasks, e.g., 15.44% mAP on VG V1.2 and 13.98% on the VG-COCO dataset.
****
Bayesian Posterior Approximation With Stochastic Ensembles
Oleksandr Balabanov, Bernhard Mehlig, Hampus Linander; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13701-13711
We introduce ensembles of stochastic neural networks to approximate the Bayesian posterior, combining stochastic methods such as dropout with deep ensembles. The stochastic ensembles are formulated as families of distributions and trained to approximate the Bayesian posterior with variational inference. We implement stochastic ensembles based on Monte Carlo dropout, DropConnect and a novel non-parametric version of dropout and evaluate them on a toy problem and CIFAR image classification. For both tasks, we test the quality of the posteriors directly against Hamiltonian Monte Carlo simulations. Our results show that stochastic ensembles provide more accurate posterior estimates than other popular baselines for Bayesian inference.
****
RILS: Masked Visual Reconstruction in Language Semantic Space
Shusheng Yang, Yixiao Ge, Kun Yi, Dian Li, Ying Shan, Xiaohu Qie, Xinggang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23304-23314
Both masked image modeling (MIM) and natural language supervision have facilitated the progress of transferable visual pre-training. In this work, we seek the synergy between two paradigms and study the emerging properties when MIM meets natural language supervision. To this end, we present a novel masked visual Reconstruction In Language semantic Space (RILS) pre-training framework, in which sentence representations, encoded by the text encoder, serve as prototypes to transform the vision-only signals into patch-sentence probabilities as semantically meaningful MIM reconstruction targets. The vision models can therefore capture useful components with structured information by predicting proper semantic of masked tokens. Better visual representations could, in turn, improve the text encode

r via the image-text alignment objective, which is essential for the effective MIM target transformation. Extensive experimental results demonstrate that our method not only enjoys the best of previous MIM and CLIP but also achieves further improvements on various tasks due to their mutual benefits. RILS exhibits advanced transferability on downstream classification, detection, and segmentation, especially for low-shot regimes. Code is available at https://github.com/hustvl/RILS.
****
Decoupling Learning and Remembering: A Bilevel Memory Framework With Knowledge Projection for Task-Incremental Learning
Wenju Sun, Qingyong Li, Jing Zhang, Wen Wang, Yangli-ao Geng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20186-20195
The dilemma between plasticity and stability arises as a common challenge for incremental learning. In contrast, the human memory system is able to remedy this dilemma owing to its multi-level memory structure, which motivates us to propose a Bilevel Memory system with Knowledge Projection (BMKP) for incremental learning. BMKP decouples the functions of learning and knowledge remembering via a bilevel-memory design: a working memory responsible for adaptively model learning, to ensure plasticity; a long-term memory in charge of enduringly storing the knowledge incorporated within the learned model, to guarantee stability. However, an emerging issue is how to extract the learned knowledge from the working memory and assimilate it into the long-term memory. To approach this issue, we reveal that the model learned by the working memory are actually residing in a redundant high-dimensional space, and the knowledge incorporated in the model can have a quite compact representation under a group of pattern basis shared by all incremental learning tasks. Therefore, we propose a knowledge projection process to adaptively maintain the shared basis, with which the loosely organized model knowledge of working memory is projected into the compact representation to be remembered in the long-term memory. We evaluate BMKP on CIFAR-10, CIFAR-100, and Tiny-ImageNet. The experimental results show that BMKP achieves state-of-the-art performance with lower memory usage.
****
R2Former: Unified Retrieval and Reranking Transformer for Place Recognition
Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, Heng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19370-19380
Visual Place Recognition (VPR) estimates the location of query images by matching them with images in a reference database. Conventional methods generally adopt aggregated CNN features for global retrieval and RANSAC-based geometric verification for reranking. However, RANSAC only employs geometric information but ignores other possible information that could be useful for reranking, e.g. local feature correlations, and attention values. In this paper, we propose a unified place recognition framework that handles both retrieval and reranking with a novel transformer model, named R2Former. The proposed reranking module takes feature correlation, attention value, and xy coordinates into account, and learns to determine whether the image pair is from the same location. The whole pipeline is end-to-end trainable and the reranking module alone can also be adopted on other CNN or transformer backbones as a generic component. Remarkably, R2Former significantly outperforms state-of-the-art methods on major VPR datasets with much less inference time and memory consumption. It also achieves the state-of-the-art on the hold-out MSLS challenge set and could serve as a simple yet strong solution for real-world large-scale applications. Experiments also show vision transformer tokens are comparable and sometimes better than CNN local features on local matching. The code is released at https://github.com/Jeff-Zilence/R2Former.
****
RepMode: Learning to Re-Parameterize Diverse Experts for Subcellular Structure Prediction
Donghao Zhou, Chunbin Gu, Junde Xu, Furui Liu, Qiong Wang, Guangyong Chen, Pheng-Ann Heng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern

Recognition (CVPR), 2023, pp. 3312-3322
In biological research, fluorescence staining is a key technique to reveal the locations and morphology of subcellular structures. However, it is slow, expensive, and harmful to cells. In this paper, we model it as a deep learning task termed subcellular structure prediction (SSP), aiming to predict the 3D fluorescent images of multiple subcellular structures from a 3D transmitted-light image. Unfortunately, due to the limitations of current biotechnology, each image is partially labeled in SSP. Besides, naturally, subcellular structures vary considerably in size, which causes the multi-scale issue of SSP. To overcome these challenges, we propose Re-parameterizing Mixture-of-Diverse-Experts (RepMode), a network that dynamically organizes its parameters with task-aware priors to handle specified single-label prediction tasks. In RepMode, the Mixture-of-Diverse-Experts (MoDE) block is designed to learn the generalized parameters for all tasks, and gating re-parameterization (GatRep) is performed to generate the specialized parameters for each task, by which RepMode can maintain a compact practical topology exactly like a plain network, and meanwhile achieves a powerful theoretical topology. Comprehensive experiments show that RepMode can achieve state-of-the-art overall performance in SSP.
****
Symmetric Shape-Preserving Autoencoder for Unsupervised Real Scene Point Cloud Completion
Changfeng Ma, Yinuo Chen, Pengxiao Guo, Jie Guo, Chongjun Wang, Yanwen Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13560-13569
Unsupervised completion of real scene objects is of vital importance but still remains extremely challenging in preserving input shapes, predicting accurate results, and adapting to multi-category data. To solve these problems, we propose in this paper an Unsupervised Symmetric Shape-Preserving Autoencoding Network, termed USSPA, to predict complete point clouds of objects from real scenes. One of our main observations is that many natural and man-made objects exhibit significant symmetries. To accommodate this, we devise a symmetry learning module to learn from those objects and to preserve structural symmetries. Starting from an initial coarse predictor, our autoencoder refines the complete shape with a carefully designed upsampling refinement module. Besides the discriminative process on the latent space, the discriminators of our USSPA also take predicted point clouds as direct guidance, enabling more detailed shape prediction. Clearly different from previous methods which train each category separately, our USSPA can be adapted to the training of multi-category data in one pass through a classifier-guided discriminator, with consistent performance on single category. For more accurate evaluation, we contribute to the community a real scene dataset with paired CAD models as ground truth. Extensive experiments and comparisons demonstrate our superiority and generalization and show that our method achieves state-of-the-art performance on unsupervised completion of real scene objects.
****
Modality-Agnostic Debiasing for Single Domain Generalization
Sanqing Qu, Yingwei Pan, Guang Chen, Ting Yao, Changjun Jiang, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24142-24151
Deep neural networks (DNNs) usually fail to generalize well to outside of distribution (OOD) data, especially in the extreme case of single domain generalization (single-DG) that transfers DNNs from single domain to multiple unseen domains. Existing single-DG techniques commonly devise various data-augmentation algorithms, and remould the multi-source domain generalization methodology to learn domain-generalized (semantic) features. Nevertheless, these methods are typically modality-specific, thereby being only applicable to one single modality (e.g., image). In contrast, we target a versatile Modality-Agnostic Debiasing (MAD) framework for single-DG, that enables generalization for different modalities. Technically, MAD introduces a novel two-branch classifier: a biased-branch encourages the classifier to identify the domain-specific (superficial) features, and a general-branch captures domain-generalized features based on the knowledge from bia

sed-branch. Our MAD is appealing in view that it is pluggable to most single-DG models. We validate the superiority of our MAD in a variety of single-DG scenarios with different modalities, including recognition on 1D texts, 2D images, 3D point clouds, and semantic segmentation on 2D images. More remarkably, for recognition on 3D point clouds and semantic segmentation on 2D images, MAD improves DSU by 2.82% and 1.5% in accuracy and mIOU.
****
Difficulty-Based Sampling for Debiased Contrastive Representation Learning
Taeuk Jang, Xiaoqian Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24039-24048
Contrastive learning is a self-supervised representation learning method that achieves milestone performance in various classification tasks. However, due to its unsupervised fashion, it suffers from the false negative sample problem: randomly drawn negative samples that are assumed to have a different label but actually have the same label as the anchor. This deteriorates the performance of contrastive learning as it contradicts the motivation of contrasting semantically similar and dissimilar pairs. This raised the attention and the importance of finding legitimate negative samples, which should be addressed by distinguishing between 1) true vs. false negatives; 2) easy vs. hard negatives. However, previous works were limited to the statistical approach to handle false negative and hard negative samples with hyperparameters tuning. In this paper, we go beyond the statistical approach and explore the connection between hard negative samples and data bias. We introduce a novel debiased contrastive learning method to explore hard negatives by relative difficulty referencing the bias-amplifying counterpart. We propose triplet loss for training a biased encoder that focuses more on easy negative samples. We theoretically show that the triplet loss amplifies the bias in self-supervised representation learning. Finally, we empirically show the proposed method improves downstream classification performance.
****
Masked Motion Encoding for Self-Supervised Video Representation Learning
Xinyu Sun, Peihao Chen, Liangwei Chen, Changhao Li, Thomas H. Li, Mingkui Tan, Chuang Gan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2235-2245
How to learn discriminative video representation from unlabeled videos is challenging but crucial for video analysis. The latest attempts seek to learn a representation model by predicting the appearance contents in the masked regions. However, simply masking and recovering appearance contents may not be sufficient to model temporal clues as the appearance contents can be easily reconstructed from a single frame. To overcome this limitation, we present Masked Motion Encoding (MME), a new pre-training paradigm that reconstructs both appearance and motion information to explore temporal clues. In MME, we focus on addressing two critical challenges to improve the representation performance: 1) how to well represent the possible long-term motion across multiple frames; and 2) how to obtain fine-grained temporal clues from sparsely sampled videos. Motivated by the fact that human is able to recognize an action by tracking objects' position changes and shape changes, we propose to reconstruct a motion trajectory that represents these two kinds of change in the masked regions. Besides, given the sparse video input, we enforce the model to reconstruct dense motion trajectories in both spatial and temporal dimensions. Pre-trained with our MME paradigm, the model is able to anticipate long-term and fine-grained motion details. Code is available at https://github.com/XinyuSun/MME.
****
CompletionFormer: Depth Completion With Convolutions and Vision Transformers
Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, Stefano Mattoccia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18527-18536
Given sparse depths and the corresponding RGB images, depth completion aims at spatially propagating the sparse measurements throughout the whole image to get a dense depth prediction. Despite the tremendous progress of deep-learning-based depth completion methods, the locality of the convolutional layer or graph model

makes it hard for the network to model the long-range relationship between pixels. While recent fully Transformer-based architecture has reported encouraging results with the global receptive field, the performance and efficiency gaps to the well-developed CNN models still exist because of its deteriorative local feature details. This paper proposes a joint convolutional attention and Transformer block (JCAT), which deeply couples the convolutional attention layer and Vision Transformer into one block, as the basic unit to construct our depth completion model in a pyramidal structure. This hybrid architecture naturally benefits both the local connectivity of convolutions and the global context of the Transformer in one single model. As a result, our CompletionFormer outperforms state-of-the-art CNNs-based methods on the outdoor KITTI Depth Completion benchmark and indoor NYUv2 dataset, achieving significantly higher efficiency (nearly 1/3 FLOPs) compared to pure Transformer-based methods. Especially when the captured depth is highly sparse, the performance gap with other methods gets much larger.
****
Comprehensive and Delicate: An Efficient Transformer for Image Restoration
Haiyu Zhao, Yuanbiao Gou, Boyun Li, Dezhong Peng, Jiancheng Lv, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14122-14132
Vision Transformers have shown promising performance in image restoration, which usually conduct window- or channel-based attention to avoid intensive computations. Although the promising performance has been achieved, they go against the biggest success factor of Transformers to a certain extent by capturing the local instead of global dependency among pixels. In this paper, we propose a novel efficient image restoration Transformer that first captures the superpixel-wise global dependency, and then transfers it into each pixel. Such a coarse-to-fine paradigm is implemented through two neural blocks, i.e., condensed attention neural block (CA) and dual adaptive neural block (DA). In brief, CA employs feature aggregation, attention computation, and feature recovery to efficiently capture the global dependency at the superpixel level. To embrace the pixel-wise global dependency, DA takes a novel dual-way structure to adaptively encapsulate the globality from superpixels into pixels. Thanks to the two neural blocks, our method achieves comparable performance while taking only 6% FLOPs compared with SwinIR.
****
Zero-Shot Model Diagnosis
Jinqi Luo, Zhaoning Wang, Chen Henry Wu, Dong Huang, Fernando De la Torre; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11631-11640
When it comes to deploying deep vision models, the behavior of these systems must be explicable to ensure confidence in their reliability and fairness. A common approach to evaluate deep learning models is to build a labeled test set with attributes of interest and assess how well it performs. However, creating a balanced test set (i.e., one that is uniformly sampled over all the important traits) is often time-consuming, expensive, and prone to mistakes. The question we try to address is: can we evaluate the sensitivity of deep learning models to arbitrary visual attributes without an annotated test set? This paper argues the case that Zero-shot Model Diagnosis (ZOOM) is possible without the need for a test set nor labeling. To avoid the need for test sets, our system relies on a generative model and CLIP. The key idea is enabling the user to select a set of prompts (relevant to the problem) and our system will automatically search for semantic counterfactual images (i.e., synthesized images that flip the prediction in the case of a binary classifier) using the generative model. We evaluate several visual tasks (classification, key-point detection, and segmentation) in multiple visual domains to demonstrate the viability of our methodology. Extensive experiments demonstrate that our method is capable of producing counterfactual images and offering sensitivity analysis for model diagnosis without the need for a test set.
****
Improving Visual Grounding by Encouraging Consistent Gradient-Based Explanations

Ziyan Yang, Kushal Kafle, Franck Dernoncourt, Vicente Ordonez; Proceedings of th
e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p
p. 19165-19174
We propose a margin-based loss for tuning joint vision-language models so that t
heir gradient-based explanations are consistent with region-level annotations pr
ovided by humans for relatively smaller grounding datasets. We refer to this obj
ective as Attention Mask Consistency (AMC) and demonstrate that it produces supe
rior visual grounding results than previous methods that rely on using vision-la
nguage models to score the outputs of object detectors. Particularly, a model tr
ained with AMC on top of standard vision-language modeling objectives obtains a
state-of-the-art accuracy of 86.49% in the Flickr30k visual grounding benchmark,
 an absolute improvement of 5.38% when compared to the best previous model train
ed under the same level of supervision. Our approach also performs exceedingly w
ell on established benchmarks for referring expression comprehension where it ob
tains 80.34% accuracy in the easy test of RefCOCO+, and 64.55% in the difficult
split. AMC is effective, easy to implement, and is general as it can be adopted
by any vision-language model, and can use any type of region annotations.
****
Physically Realizable Natural-Looking Clothing Textures Evade Person Detectors v
ia 3D Modeling
Zhanhao Hu, Wenda Chu, Xiaopei Zhu, Hui Zhang, Bo Zhang, Xiaolin Hu; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 16975-16984
Recent works have proposed to craft adversarial clothes for evading person detec
tors, while they are either only effective at limited viewing angles or very con
spicuous to humans. We aim to craft adversarial texture for clothes based on 3D
modeling, an idea that has been used to craft rigid adversarial objects such as
a 3D-printed turtle. Unlike rigid objects, humans and clothes are non-rigid, lea
ding to difficulties in physical realization. In order to craft natural-looking
adversarial clothes that can evade person detectors at multiple viewing angles,
we propose adversarial camouflage textures (AdvCaT) that resemble one kind of th
e typical textures of daily clothes, camouflage textures. We leverage the Vorono
i diagram and Gumbel-softmax trick to parameterize the camouflage textures and o
ptimize the parameters via 3D modeling. Moreover, we propose an efficient augmen
tation pipeline on 3D meshes combining topologically plausible projection (TopoP
roj) and Thin Plate Spline (TPS) to narrow the gap between digital and real-worl
d objects. We printed the developed 3D texture pieces on fabric materials and ta
ilored them into T-shirts and trousers. Experiments show high attack success rat
es of these clothes against multiple detectors.
****
ShadowDiffusion: When Degradation Prior Meets Diffusion Model for Shadow Removal
Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister,
 Bihan Wen; Proceedings of the IEEE/CVF Conference on Computer Vision and Patter
n Recognition (CVPR), 2023, pp. 14049-14058
Recent deep learning methods have achieved promising results in image shadow rem
oval. However, their restored images still suffer from unsatisfactory boundary a
rtifacts, due to the lack of degradation prior and the deficiency in modeling ca
pacity. Our work addresses these issues by proposing a unified diffusion framewo
rk that integrates both the image and degradation priors for highly effective sh
adow removal. In detail, we first propose a shadow degradation model, which insp
ires us to build a novel unrolling diffusion model, dubbed ShandowDiffusion. It
remarkably improves the model's capacity in shadow removal via progressively ref
ining the desired output with both degradation prior and diffusive generative pr
ior, which by nature can serve as a new strong baseline for image restoration. F
urthermore, ShadowDiffusion progressively refines the estimated shadow mask as a
n auxiliary task of the diffusion generator, which leads to more accurate and ro
bust shadow-free image generation. We conduct extensive experiments on three pop
ular public datasets, including ISTD, ISTD+, and SRD, to validate our method's e
ffectiveness. Compared to the state-of-the-art methods, our model achieves a sig
nificant improvement in terms of PSNR, increasing from 31.69dB to 34.73dB over S

RD dataset.
****
FFHQ-UV: Normalized Facial UV-Texture Dataset for 3D Face Reconstruction
Haoran Bai, Di Kang, Haoxian Zhang, Jinshan Pan, Linchao Bao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 362-371
We present a large-scale facial UV-texture dataset that contains over 50,000 high-quality texture UV-maps with even illuminations, neutral expressions, and cleaned facial regions, which are desired characteristics for rendering realistic 3D face models under different lighting conditions. The dataset is derived from a large-scale face image dataset namely FFHQ, with the help of our fully automatic and robust UV-texture production pipeline. Our pipeline utilizes the recent advances in StyleGAN-based facial image editing approaches to generate multi-view normalized face images from single-image inputs. An elaborated UV-texture extraction, correction, and completion procedure is then applied to produce high-quality UV-maps from the normalized face images. Compared with existing UV-texture datasets, our dataset has more diverse and higher-quality texture maps. We further train a GAN-based texture decoder as the nonlinear texture basis for parametric fitting based 3D face reconstruction. Experiments show that our method improves the reconstruction accuracy over state-of-the-art approaches, and more importantly, produces high-quality texture maps that are ready for realistic renderings. The dataset, code, and pre-trained texture decoder are publicly available at https://github.com/csbhr/FFHQ-UV.
****
Pruning Parameterization With Bi-Level Optimization for Efficient Semantic Segmentation on the Edge
Changdi Yang, Pu Zhao, Yanyu Li, Wei Niu, Jiexiong Guan, Hao Tang, Minghai Qin, Bin Ren, Xue Lin, Yanzhi Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15402-15412
With the ever-increasing popularity of edge devices, it is necessary to implement real-time segmentation on the edge for autonomous driving and many other applications. Vision Transformers (ViTs) have shown considerably stronger results for many vision tasks. However, ViTs with the full-attention mechanism usually consume a large number of computational resources, leading to difficulties for real-time inference on edge devices. In this paper, we aim to derive ViTs with fewer computations and fast inference speed to facilitate the dense prediction of semantic segmentation on edge devices. To achieve this, we propose a pruning parameterization method to formulate the pruning problem of semantic segmentation. Then we adopt a bi-level optimization method to solve this problem with the help of implicit gradients. Our experimental results demonstrate that we can achieve 38.9 mIoU on ADE20K val with a speed of 56.5 FPS on Samsung S21, which is the highest mIoU under the same computation constraint with real-time inference.
****
Camouflaged Object Detection With Feature Decomposition and Edge Reconstruction
Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, Xiu Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22046-22055
Camouflaged object detection (COD) aims to address the tough issue of identifying camouflaged objects visually blended into the surrounding backgrounds. COD is a challenging task due to the intrinsic similarity of camouflaged objects with the background, as well as their ambiguous boundaries. Existing approaches to this problem have developed various techniques to mimic the human visual system. Albeit effective in many cases, these methods still struggle when camouflaged objects are so deceptive to the vision system. In this paper, we propose the FEature Decomposition and Edge Reconstruction (FEDER) model for COD. The FEDER model addresses the intrinsic similarity of foreground and background by decomposing the features into different frequency bands using learnable wavelets. It then focuses on the most informative bands to mine subtle cues that differentiate foreground and background. To achieve this, a frequency attention module and a guidance-based feature aggregation module are developed. To combat the ambiguous boundary

 problem, we propose to learn an auxiliary edge reconstruction task alongside the COD task. We design an ordinary differential equation-inspired edge reconstruction module that generates exact edges. By learning the auxiliary task in conjunction with the COD task, the FEDER model can generate precise prediction maps with accurate object boundaries. Experiments show that our FEDER model significantly outperforms state-of-the-art methods with cheaper computational and memory costs.
****
ALOFT: A Lightweight MLP-Like Architecture With Dynamic Low-Frequency Transform for Domain Generalization
Jintao Guo, Na Wang, Lei Qi, Yinghuan Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24132-24141
Domain generalization (DG) aims to learn a model that generalizes well to unseen target domains utilizing multiple source domains without re-training. Most existing DG works are based on convolutional neural networks (CNNs). However, the local operation of the convolution kernel makes the model focus too much on local representations (e.g., texture), which inherently causes the model more prone to overfit to the source domains and hampers its generalization ability. Recently, several MLP-based methods have achieved promising results in supervised learning tasks by learning global interactions among different patches of the image. Inspired by this, in this paper, we first analyze the difference between CNN and MLP methods in DG and find that MLP methods exhibit a better generalization ability because they can better capture the global representations (e.g., structure) than CNN methods. Then, based on a recent lightweight MLP method, we obtain a strong baseline that outperforms most start-of-the-art CNN-based methods. The baseline can learn global structure representations with a filter to suppress structure-irrelevant information in the frequency space. Moreover, we propose a dynAmic LOw-Frequency spectrum Transform (ALOFT) that can perturb local texture features while preserving global structure features, thus enabling the filter to remove structure-irrelevant information sufficiently. Extensive experiments on four benchmarks have demonstrated that our method can achieve great performance improvement with a small number of parameters compared to SOTA CNN-based DG methods. Our code is available at https://github.com/lingeringlight/ALOFT/.
****
NLOST: Non-Line-of-Sight Imaging With Transformer
Yue Li, Jiayong Peng, Juntian Ye, Yueyi Zhang, Feihu Xu, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13313-13322
Time-resolved non-line-of-sight (NLOS) imaging is based on the multi-bounce indirect reflections from the hidden objects for 3D sensing. Reconstruction from NLOS measurements remains challenging especially for complicated scenes. To boost the performance, we present NLOST, the first transformer-based neural network for NLOS reconstruction. Specifically, after extracting the shallow features with the assistance of physics-based priors, we design two spatial-temporal self attention encoders to explore both local and global correlations within 3D NLOS data by splitting or downsampling the features into different scales, respectively. Then, we design a spatial-temporal cross attention decoder to integrate local and global features in the token space of transformer, resulting in deep features with high representation capabilities. Finally, deep and shallow features are fused to reconstruct the 3D volume of hidden scenes. Extensive experimental results demonstrate the superior performance of the proposed method over existing solutions on both synthetic data and real-world data captured by different NLOS imaging systems.
****
Text-Visual Prompting for Efficient 2D Temporal Video Grounding
Yimeng Zhang, Xin Chen, Jinghan Jia, Sijia Liu, Ke Ding; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14794-14804
In this paper, we study the problem of temporal video grounding (TVG), which aims to predict the starting/ending time points of moments described by a text sent

ence within a long untrimmed video. Benefiting from fine-grained 3D visual featu res, the TVG techniques have achieved remarkable progress in recent years. Howev er, the high complexity of 3D convolutional neural networks (CNNs) makes extract ing dense 3D visual features time-consuming, which calls for intensive memory an d computing resources. Towards efficient TVG, we propose a novel text-visual pro mpting (TVP) framework, which incorporates optimized perturbation patterns (that we call 'prompts') into both visual inputs and textual features of a TVG model. In sharp contrast to 3D CNNs, we show that TVP allows us to effectively co-trai n vision encoder and language encoder in a 2D TVG model and improves the perform ance of crossmodal feature fusion using only low-complexity sparse 2D visual fea tures. Further, we propose a Temporal-Distance IoU (TDIoU) loss for efficient le arning of TVG. Experiments on two benchmark datasets, Charades-STA and ActivityN et Captions datasets, empirically show that the proposed TVP significantly boost s the performance of 2D TVG (e.g., 9.79% improvement on Charades-STA and 30.77% improvement on ActivityNet Captions) and achieves 5x inference acceleration over TVG using 3D visual features. Codes are available at Open.Intel.
****
SurfelNeRF: Neural Surfel Radiance Fields for Online Photorealistic Reconstructi on of Indoor Scenes
Yiming Gao, Yan-Pei Cao, Ying Shan; Proceedings of the IEEE/CVF Conference on Co mputer Vision and Pattern Recognition (CVPR), 2023, pp. 108-118
Online reconstructing and rendering of large-scale indoor scenes is a long-stand ing challenge. SLAM-based methods can reconstruct 3D scene geometry progressivel y in real time but can not render photorealistic results. While NeRF-based metho ds produce promising novel view synthesis results, their long offline optimizati on time and lack of geometric constraints pose challenges to efficiently handlin g online input. Inspired by the complementary advantages of classical 3D reconst ruction and NeRF, we thus investigate marrying explicit geometric representation with NeRF rendering to achieve efficient online reconstruction and high-quality rendering. We introduce SurfelNeRF, a variant of neural radiance field which em ploys a flexible and scalable neural surfel representation to store geometric at tributes and extracted appearance features from input images. We further extend conventional surfel-based fusion scheme to progressively integrate incoming inpu t frames into the reconstructed global neural scene representation. In addition, we propose a highly-efficient differentiable rasterization scheme for rendering neural surfel radiance fields, which helps SurfelNeRF achieve 10x speedups in t raining and inference time, respectively. Experimental results show that our met hod achieves the state-of-the-art 23.82 PSNR and 29.58 PSNR on ScanNet in feedfo rward inference and per-scene optimization settings, respectively.
****
Learning Visual Representations via Language-Guided Sampling
Mohamed El Banani, Karan Desai, Justin Johnson; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19208-19220
Although an object may appear in numerous contexts, we often describe it in a li mited number of ways. Language allows us to abstract away visual variation to re present and communicate concepts. Building on this intuition, we propose an alte rnative approach to visual representation learning: using language similarity to sample semantically similar image pairs for contrastive learning. Our approach diverges from image-based contrastive learning by sampling view pairs using lang uage similarity instead of hand-crafted augmentations or learned clusters. Our a pproach also differs from image-text contrastive learning by relying on pre-trai ned language models to guide the learning rather than directly minimizing a cros s-modal loss. Through a series of experiments, we show that language-guided lear ning yields better features than image-based and image-text representation learn ing approaches.
****
Logical Implications for Visual Question Answering Consistency
Sergio Tascon-Morales, Pablo Márquez-Neila, Raphael Sznitman; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp . 6725-6735

Despite considerable recent progress in Visual Question Answering (VQA) models, inconsistent or contradictory answers continue to cast doubt on their true reasoning capabilities. However, most proposed methods use indirect strategies or strong assumptions on pairs of questions and answers to enforce model consistency. Instead, we propose a novel strategy intended to improve model performance by directly reducing logical inconsistencies. To do this, we introduce a new consistency loss term that can be used by a wide range of the VQA models and which relies on knowing the logical relation between pairs of questions and answers. While such information is typically not available in VQA datasets, we propose to infer these logical relations using a dedicated language model and use these in our proposed consistency loss function. We conduct extensive experiments on the VQA Introspect and DME datasets and show that our method brings improvements to state-of-the-art VQA models while being robust across different architectures and settings.
****
NeUDF: Leaning Neural Unsigned Distance Fields With Volume Rendering
Yu-Tao Liu, Li Wang, Jie Yang, Weikai Chen, Xiaoxu Meng, Bo Yang, Lin Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 237-247
Multi-view shape reconstruction has achieved impressive progresses thanks to the latest advances in neural implicit surface rendering. However, existing methods based on signed distance function (SDF) are limited to closed surfaces, failing to reconstruct a wide range of real-world objects that contain open-surface structures. In this work, we introduce a new neural rendering framework, coded NeUDF, that can reconstruct surfaces with arbitrary topologies solely from multi-view supervision. To gain the flexibility of representing arbitrary surfaces, NeUDF leverages the unsigned distance function (UDF) as surface representation. While a naive extension of SDF-based neural renderer cannot scale to UDF, we propose two new formulations of weight function specially tailored for UDF-based volume rendering. Furthermore, to cope with open surface rendering, where the in/out test is no longer valid, we present a dedicated normal regularization strategy to resolve the surface orientation ambiguity. We extensively evaluate our method over a number of challenging datasets, including DTU, MGN, and Deep Fashion 3D. Experimental results demonstrate that NeUDF can significantly outperform the state-of-the-art method in the task of multi-view surface reconstruction, especially for the complex shapes with open boundaries.
****
Master: Meta Style Transformer for Controllable Zero-Shot and Few-Shot Artistic Style Transfer
Hao Tang, Songhua Liu, Tianwei Lin, Shaoli Huang, Fu Li, Dongliang He, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18329-18338
Transformer-based models achieve favorable performance in artistic style transfer recently thanks to its global receptive field and powerful multi-head/layer attention operations. Nevertheless, the over-paramerized multi-layer structure increases parameters significantly and thus presents a heavy burden for training. Moreover, for the task of style transfer, vanilla Transformer that fuses content and style features by residual connections is prone to content-wise distortion. In this paper, we devise a novel Transformer model termed as Master specifically for style transfer. On the one hand, in the proposed model, different Transformer layers share a common group of parameters, which (1) reduces the total number of parameters, (2) leads to more robust training convergence, and (3) is readily to control the degree of stylization via tuning the number of stacked layers freely during inference. On the other hand, different from the vanilla version, we adopt a learnable scaling operation on content features before content-style feature interaction, which better preserves the original similarity between a pair of content features while ensuring the stylization quality. We also propose a novel meta learning scheme for the proposed model so that it can not only work in the typical setting of arbitrary style transfer, but also adaptable to the few-shot setting, by only fine-tuning the Transformer encoder layer in the few-shot

stage for one specific style. Text-guided few-shot style transfer is firstly ac hieved with the proposed framework. Extensive experiments demonstrate the superi ority of Master under both zero-shot and few-shot style transfer settings.
****
Affordance Diffusion: Synthesizing Hand-Object Interactions
Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, Sifei Liu; Proceedings of the IEEE/CVF Conference on Com puter Vision and Pattern Recognition (CVPR), 2023, pp. 22479-22489
Recent successes in image synthesis are powered by large-scale diffusion models. However, most methods are currently limited to either text- or image-conditione d generation for synthesizing an entire image, texture transfer or inserting obj ects into a user-specified region. In contrast, in this work we focus on synthes izing complex interactions (i.e., an articulated hand) with a given object. Give n an RGB image of an object, we aim to hallucinate plausible images of a human h and interacting with it. We propose a two step generative approach that leverage s a LayoutNet that samples an articulation-agnostic hand-object-interaction layo ut, and a ContentNet that synthesizes images of a hand grasping the object given the predicted layout. Both are built on top of a large-scale pretrained diffusi on model to make use of its latent representation. Compared to baselines, the pr oposed method is shown to generalize better to novel objects and perform surpris ingly well on out-of-distribution in-the-wild scenes. The resulting system allow s us to predict descriptive affordance information, such as hand articulation an d approaching orientation.
****
NEF: Neural Edge Fields for 3D Parametric Curve Reconstruction From Multi-View I mages
Yunfan Ye, Renjiao Yi, Zhirui Gao, Chenyang Zhu, Zhiping Cai, Kai Xu; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8486-8495
We study the problem of reconstructing 3D feature curves of an object from a set of calibrated multi-view images. To do so, we learn a neural implicit field rep resenting the density distribution of 3D edges which we refer to as Neural Edge Field (NEF). Inspired by NeRF, NEF is optimized with a view-based rendering loss where a 2D edge map is rendered at a given view and is compared to the ground-t ruth edge map extracted from the image of that view. The rendering-based differe ntiable optimization of NEF fully exploits 2D edge detection, without needing a supervision of 3D edges, a 3D geometric operator or cross-view edge corresponden ce. Several technical designs are devised to ensure learning a range-limited and view-independent NEF for robust edge extraction. The final parametric 3D curves are extracted from NEF with an iterative optimization method. On our benchmark with synthetic data, we demonstrate that NEF outperforms existing state-of-the-a rt methods on all metrics. Project page: https://yunfan1202.github.io/NEF/.
****
Geometric Visual Similarity Learning in 3D Medical Image Self-Supervised Pre-Tra ining
Yuting He, Guanyu Yang, Rongjun Ge, Yang Chen, Jean-Louis Coatrieux, Boyu Wang, Shuo Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2023, pp. 9538-9547
Learning inter-image similarity is crucial for 3D medical images self-supervised pre-training, due to their sharing of numerous same semantic regions. However, the lack of the semantic prior in metrics and the semantic-independent variation in 3D medical images make it challenging to get a reliable measurement for the inter-image similarity, hindering the learning of consistent representation for same semantics. We investigate the challenging problem of this task, i.e., learn ing a consistent representation between images for a clustering effect of same s emantic features. We propose a novel visual similarity learning paradigm, Geomet ric Visual Similarity Learning, which embeds the prior of topological invariance into the measurement of the inter-image similarity for consistent representatio n of semantic regions. To drive this paradigm, we further construct a novel geom etric matching head, the Z-matching head, to collaboratively learn the global an

d local similarity of semantic regions, guiding the efficient representation lea
rning for different scale-level inter-image semantic features. Our experiments d
emonstrate that the pre-training with our learning of inter-image similarity yie
lds more powerful inner-scene, inter-scene, and global-local transferring abilit
y on four challenging 3D medical image tasks. Our codes and pre-trained models w
ill be publicly available in https://github.com/YutingHe-list/GVSL.
****
Towards Artistic Image Aesthetics Assessment: A Large-Scale Dataset and a New Me
thod
Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, Paul L. Rosin; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
22388-22397
Image aesthetics assessment (IAA) is a challenging task due to its highly subjec
tive nature. Most of the current studies rely on large-scale datasets (e.g., AVA
 and AADB) to learn a general model for all kinds of photography images. However
, little light has been shed on measuring the aesthetic quality of artistic imag
es, and the existing datasets only contain relatively few artworks. Such a defec
t is a great obstacle to the aesthetic assessment of artistic images. To fill th
e gap in the field of artistic image aesthetics assessment (AIAA), we first intr
oduce a large-scale AIAA dataset: Boldbrush Artistic Image Dataset (BAID), which
 consists of 60,337 artistic images covering various art forms, with more than 3
60,000 votes from online users. We then propose a new method, SAAN (Style-specif
ic Art Assessment Network), which can effectively extract and utilize style-spec
ific and generic aesthetic information to evaluate artistic images. Experiments
demonstrate that our proposed approach outperforms existing IAA methods on the p
roposed BAID dataset according to quantitative comparisons. We believe the propo
sed dataset and method can serve as a foundation for future AIAA works and inspi
re more research in this field.
****
MM-3DScene: 3D Scene Understanding by Customizing Masked Modeling With Informati
ve-Preserved Reconstruction and Self-Distilled Consistency
Mingye Xu, Mutian Xu, Tong He, Wanli Ouyang, Yali Wang, Xiaoguang Han, Yu Qiao;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2023, pp. 4380-4390
Masked Modeling (MM) has demonstrated widespread success in various vision chall
enges, by reconstructing masked visual patches. Yet, applying MM for large-scale
 3D scenes remains an open problem due to the data sparsity and scene complexity
. The conventional random masking paradigm used in 2D images often causes a high
 risk of ambiguity when recovering the masked region of 3D scenes. To this end,
we propose a novel informative-preserved reconstruction, which explores local st
atistics to discover and preserve the representative structured points, effectiv
ely enhancing the pretext masking task for 3D scene understanding. Integrated wi
th a progressive reconstruction manner, our method can concentrate on modeling r
egional geometry and enjoy less ambiguity for masked reconstruction. Besides, su
ch scenes with progressive masking ratios can also serve to self-distill their i
ntrinsic spatial consistency, requiring to learn the consistent representations
from unmasked areas. By elegantly combining informative-preserved reconstruction
 on masked areas and consistency self-distillation from unmasked areas, a unifie
d framework called MM-3DScene is yielded. We conduct comprehensive experiments o
n a host of downstream tasks. The consistent improvement (e.g., +6.1% mAP@0.5 on
 object detection and +2.2% mIoU on semantic segmentation) demonstrates the supe
riority of our approach.
****
Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation
Narek Tumanyan, Michal Geyer, Shai Bagon, Tali Dekel; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1921-1
930
Large-scale text-to-image generative models have been a revolutionary breakthrou
gh in the evolution of generative AI, synthesizing diverse images with highly co
mplex visual concepts. However, a pivotal challenge in leveraging such models fo

r real-world content creation is providing users with control over the generated
 content. In this paper, we present a new framework that takes text-to-image syn
thesis to the realm of image-to-image translation -- given a guidance image and
a target text prompt as input, our method harnesses the power of a pre-trained t
ext-to-image diffusion model to generate a new image that complies with the targ
et text, while preserving the semantic layout of the guidance image. Specificall
y, we observe and empirically demonstrate that fine-grained control over the gen
erated structure can be achieved by manipulating spatial features and their self
-attention inside the model. This results in a simple and effective approach, wh
ere features extracted from the guidance image are directly injected into the ge
neration process of the translated image, requiring no training or fine-tuning.
We demonstrate high-quality results on versatile text-guided image translation t
asks, including translating sketches, rough drawings and animations into realist
ic images, changing the class and appearance of objects in a given image, and mo
difying global qualities such as lighting and color.
****
Inverting the Imaging Process by Learning an Implicit Camera Model
Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Qing Wang; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2145
6-21465
Representing visual signals with implicit coordinate-based neural networks, as a
n effective replacement of the traditional discrete signal representation, has g
ained considerable popularity in computer vision and graphics. In contrast to ex
isting implicit neural representations which focus on modelling the scene only,
this paper proposes a novel implicit camera model which represents the physical
imaging process of a camera as a deep neural network. We demonstrate the power o
f this new implicit camera model on two inverse imaging tasks: i) generating all
-in-focus photos, and ii) HDR imaging. Specifically, we devise an implicit blur
generator and an implicit tone mapper to model the aperture and exposure of the
camera's imaging process, respectively. Our implicit camera model is jointly lea
rned together with implicit scene models under multi-focus stack and multi-expos
ure bracket supervision. We have demonstrated the effectiveness of our new model
 on large number of test images and videos, producing accurate and visually appe
aling all-in-focus and high dynamic range images. In principle, our new implicit
 neural camera model has the potential to benefit a wide array of other inverse
imaging tasks.
****
Fast Contextual Scene Graph Generation With Unbiased Context Augmentation
Tianlei Jin, Fangtai Guo, Qiwei Meng, Shiqiang Zhu, Xiangming Xi, Wen Wang, Zong
hao Mu, Wei Song; Proceedings of the IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR), 2023, pp. 6302-6311
Scene graph generation (SGG) methods have historically suffered from long-tail b
ias and slow inference speed. In this paper, we notice that humans can analyze r
elationships between objects relying solely on context descriptions,and this abs
tract cognitive process may be guided by experience. For example, given descript
ions of cup and table with their spatial locations, humans can speculate possibl
e relationships < cup, on, table > or < table, near, cup >. Even without visual
appearance information, some impossible predicates like flying in and looking at
 can be empirically excluded. Accordingly, we propose a contextual scene graph g
eneration (C-SGG) method without using visual information and introduce a contex
t augmentation method. We propose that slight perturbations in the position and
size of objects do not essentially affect the relationship between objects. Ther
efore, at the context level, we can produce diverse context descriptions by usin
g a context augmentation method based on the original dataset. These diverse con
text descriptions can be used for unbiased training of C-SGG to alleviate long-t
ail bias. In addition, we also introduce a context guided visual scene graph gen
eration (CV-SGG) method, which leverages the C-SGG experience to guide vision to
 focus on possible predicates. Through extensive experiments on the publicly ava
ilable dataset, C-SGG alleviates long-tail bias and omits the huge computation o
f visual feature extraction to realize real-time SGG. CV-SGG achieves a great tr

ade-off between common predicates and tail predicates.
****
Less Is More: Reducing Task and Model Complexity for 3D Point Cloud Semantic Seg mentation
Li Li, Hubert P. H. Shum, Toby P. Breckon; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9361-9371
Whilst the availability of 3D LiDAR point cloud data has significantly grown in recent years, annotation remains expensive and time-consuming, leading to a dema nd for semi-supervised semantic segmentation methods with application domains su ch as autonomous driving. Existing work very often employs relatively large segm entation backbone networks to improve segmentation accuracy, at the expense of c omputational costs. In addition, many use uniform sampling to reduce ground trut h data requirements for learning needed, often resulting in sub-optimal performa nce. To address these issues, we propose a new pipeline that employs a smaller a rchitecture, requiring fewer ground-truth annotations to achieve superior segmen tation accuracy compared to contemporary approaches. This is facilitated via a n ovel Sparse Depthwise Separable Convolution module that significantly reduces th e network parameter count while retaining overall task performance. To effective ly sub-sample our training data, we propose a new Spatio-Temporal Redundant Fram e Downsampling (ST-RFD) method that leverages knowledge of sensor motion within the environment to extract a more diverse subset of training data frame samples. To leverage the use of limited annotated data samples, we further propose a sof t pseudo-label method informed by LiDAR reflectivity. Our method outperforms con temporary semi-supervised work in terms of mIoU, using less labeled data, on the SemanticKITTI (59.5@5%) and ScribbleKITTI (58.1@5%) benchmark datasets, based o n a 2.3x reduction in model parameters and 641x fewer multiply-add operations wh ilst also demonstrating significant performance improvement on limited training data (i.e., Less is More).
****
Re-Thinking Federated Active Learning Based on Inter-Class Diversity
SangMook Kim, Sangmin Bae, Hwanjun Song, Se-Young Yun; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3944-3953
Although federated learning has made awe-inspiring advances, most studies have a ssumed that the client's data are fully labeled. However, in a real-world scenar io, every client may have a significant amount of unlabeled instances. Among the various approaches to utilizing unlabeled data, a federated active learning fra mework has emerged as a promising solution. In the decentralized setting, there are two types of available query selector models, namely 'global' and 'local-onl y' models, but little literature discusses their performance dominance and its c auses. In this work, we first demonstrate that the superiority of two selector m odels depends on the global and local inter-class diversity. Furthermore, we obs erve that the global and local-only models are the keys to resolving the imbalan ce of each side. Based on our findings, we propose LoGo, a FAL sampling strategy robust to varying local heterogeneity levels and global imbalance ratio, that i ntegrates both models by two steps of active selection scheme. LoGo consistently outperforms six active learning strategies in the total number of 38 experiment al settings.
****
Enhanced Training of Query-Based Object Detection via Selective Query Recollecti on
Fangyi Chen, Han Zhang, Kai Hu, Yu-Kai Huang, Chenchen Zhu, Marios Savvides; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23756-23765
This paper investigates a phenomenon where query-based object detectors mispredi ct at the last decoding stage while predicting correctly at an intermediate stag e. We review the training process and attribute the overlooked phenomenon to two limitations: lack of training emphasis and cascading errors from decoding seque nce. We design and present Selective Query Recollection (SQR), a simple and effe ctive training strategy for query-based object detectors. It cumulatively collec

ts intermediate queries as decoding stages go deeper and selectively forwards th
e queries to the downstream stages aside from the sequential structure. Such-wis
e, SQR places training emphasis on later stages and allows later stages to work
with intermediate queries from earlier stages directly. SQR can be easily plugge
d into various query-based object detectors and significantly enhances their per
formance while leaving the inference pipeline unchanged. As a result, we apply S
QR on Adamixer, DAB-DETR, and Deformable-DETR across various settings (backbone,
 number of queries, schedule) and consistently brings 1.4   2.8 AP improvement.
****
AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning With Masked Autoe
ncoders
Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal A
grawal, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Visi
on and Pattern Recognition (CVPR), 2023, pp. 14507-14517
Masked Autoencoders (MAEs) learn generalizable representations for image, text,
audio, video, etc., by reconstructing masked input data from tokens of the visib
le data. Current MAE approaches for videos rely on random patch, tube, or frame
based masking strategies to select these tokens. This paper proposes AdaMAE, an
adaptive masking strategy for MAEs that is end-to-end trainable. Our adaptive ma
sking strategy samples visible tokens based on the semantic context using an aux
iliary sampling network. This network estimates a categorical distribution over
spacetime-patch tokens. The tokens that increase the expected reconstruction err
or are rewarded and selected as visible tokens, motivated by the policy gradient
 algorithm in reinforcement learning. We show that AdaMAE samples more tokens fr
om the high spatiotemporal information regions, thereby allowing us to mask 95%
of tokens, resulting in lower memory requirements and faster pre-training. We co
nduct ablation studies on the Something-Something v2 (SSv2) dataset to demonstra
te the efficacy of our adaptive sampling approach and report state-of-the-art re
sults of 70.0% and 81.7% in top-1 accuracy on SSv2 and Kinetics-400 action class
ification datasets with a ViT-Base backbone and 800 pre-training epochs. Code an
d pre-trained models are available at: https://github.com/wgcban/adamae.git
****
Detecting Human-Object Contact in Images
Yixin Chen, Sai Kumar Dwivedi, Michael J. Black, Dimitrios Tzionas; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
23, pp. 17100-17110
Humans constantly contact objects to move and perform tasks. Thus, detecting hum
an-object contact is important for building human-centered artificial intelligen
ce. However, there exists no robust method to detect contact between the body an
d the scene from an image, and there exists no dataset to learn such a detector.
 We fill this gap with HOT ("Human-Object conTact"), a new dataset of human-obje
ct contacts in images. To build HOT, we use two data sources: (1) We use the PRO
X dataset of 3D human meshes moving in 3D scenes, and automatically annotate 2D
image areas for contact via 3D mesh proximity and projection. (2) We use the V-C
OCO, HAKE and Watch-n-Patch datasets, and ask trained annotators to draw polygon
s around the 2D image areas where contact takes place. We also annotate the invo
lved body part of the human body. We use our HOT dataset to train a new contact
detector, which takes a single color image as input, and outputs 2D contact heat
maps as well as the body-part labels that are in contact. This is a new and chal
lenging task, that extends current foot-ground or hand-object contact detectors
to the full generality of the whole body. The detector uses a part-attention bra
nch to guide contact estimation through the context of the surrounding body part
s and scene. We evaluate our detector extensively, and quantitative results show
 that our model outperforms baselines, and that all components contribute to bet
ter performance. Results on images from an online repository show reasonable det
ections and generalizability. Our HOT data and model are available for research
at https://hot.is.tue.mpg.de.
****
PointClustering: Unsupervised Point Cloud Pre-Training Using Transformation Inva
riance in Clustering

Fuchen Long, Ting Yao, Zhaofan Qiu, Lusong Li, Tao Mei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21824-21834
Feature invariance under different data transformations, i.e., transformation invariance, can be regarded as a type of self-supervision for representation learning. In this paper, we present PointClustering, a new unsupervised representation learning scheme that leverages transformation invariance for point cloud pre-training. PointClustering formulates the pretext task as deep clustering and employs transformation invariance as an inductive bias, following the philosophy that common point cloud transformation will not change the geometric properties and semantics. Technically, PointClustering iteratively optimizes the feature clusters and backbone, and delves into the transformation invariance as learning regularization from two perspectives: point level and instance level. Point-level invariance learning maintains local geometric properties through gathering point features of one instance across transformations, while instance-level invariance learning further measures clusters over the entire dataset to explore semantics of instances. Our PointClustering is architecture-agnostic and readily applicable to MLP-based, CNN-based and Transformer-based backbones. We empirically demonstrate that the models pre-learnt on the ScanNet dataset by PointClustering provide superior performances on six benchmarks, across downstream tasks of classification and segmentation. More remarkably, PointClustering achieves an accuracy of 94.5% on ModelNet40 with Transformer backbone. Source code is available at https://github.com/FuchenUSTC/PointClustering.
****
CiaoSR: Continuous Implicit Attention-in-Attention Network for Arbitrary-Scale Image Super-Resolution
Jiezhang Cao, Qin Wang, Yongqin Xian, Yawei Li, Bingbing Ni, Zhiming Pi, Kai Zhang, Yulun Zhang, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1796-1807
Learning continuous image representations is recently gaining popularity for image super-resolution (SR) because of its ability to reconstruct high-resolution images with arbitrary scales from low-resolution inputs. Existing methods mostly ensemble nearby features to predict the new pixel at any queried coordinate in the SR image. Such a local ensemble suffers from some limitations: i) it has no learnable parameters and it neglects the similarity of the visual features; ii) it has a limited receptive field and cannot ensemble relevant features in a large field which are important in an image. To address these issues, this paper proposes a continuous implicit attention-in-attention network, called CiaoSR. We explicitly design an implicit attention network to learn the ensemble weights for the nearby local features. Furthermore, we embed a scale-aware attention in this implicit attention network to exploit additional non-local information. Extensive experiments on benchmark datasets demonstrate CiaoSR significantly outperforms the existing single image SR methods with the same backbone. In addition, CiaoSR also achieves the state-of-the-art performance on the arbitrary-scale SR task. The effectiveness of the method is also demonstrated on the real-world SR setting. More importantly, CiaoSR can be flexibly integrated into any backbone to improve the SR performance.
****
Out-of-Distributed Semantic Pruning for Robust Semi-Supervised Learning
Yu Wang, Pengchong Qiao, Chang Liu, Guoli Song, Xiawu Zheng, Jie Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23849-23858
Recent advances in robust semi-supervised learning (SSL) typical filters out-of-distribution (OOD) information at the sample level. We argue that an overlooked problem of robust SSL is its corrupted information on semantic level, practically limiting the development of the field. In this paper, we take an initiative step to explore and propose a unified framework termed as OOD Semantic Pruning (OSP), aims at pruning OOD semantics out from the in-distribution (ID) features. Specifically, (i) we propose an aliasing OOD matching module to pair each ID sample with an OOD sample with semantic overlap. (ii) We design a soft orthogonality

regularization, which first transforms each ID feature by suppressing its semant
ic component that is collinear with paired OOD sample. It then forces the predic
tions before and after soft orthogonality transformation to be consistent. Being
 practically simple, our method shows a strong performance in OOD detection and
ID classification on challenging benchmarks. In particular, OSP surpasses the pr
evious state-of-the-art by 13.7% on accuracy for ID classification and 5.9% on A
UROC for OOD detection on TinyImageNet dataset. Codes are available in the suppl
ementary material.
****
The Best Defense Is a Good Offense: Adversarial Augmentation Against Adversarial
 Attacks
Iuri Frosio, Jan Kautz; Proceedings of the IEEE/CVF Conference on Computer Visio
n and Pattern Recognition (CVPR), 2023, pp. 4067-4076
Many defenses against adversarial attacks (e.g. robust classifiers, randomizatio
n, or image purification) use countermeasures put to work only after the attack
has been crafted. We adopt a different perspective to introduce A^5 (Adversarial
 Augmentation Against Adversarial Attacks), a novel framework including the firs
t certified preemptive defense against adversarial attacks. The main idea is to
craft a defensive perturbation to guarantee that any attack (up to a given magni
tude) towards the input in hand will fail. To this aim, we leverage existing aut
omatic perturbation analysis tools for neural networks. We study the conditions
to apply A^5 effectively, analyze the importance of the robustness of the to-be-
defended classifier, and inspect the appearance of the robustified images. We sh
ow effective on-the-fly defensive augmentation with a robustifier network that i
gnores the ground truth label, and demonstrate the benefits of robustifier and c
lassifier co-training. In our tests, A^5 consistently beats state of the art cer
tified defenses on MNIST, CIFAR10, FashionMNIST and Tinyimagenet. We also show h
ow to apply A^5 to create certifiably robust physical objects. The released code
 at https://github.com/NVlabs/A5 allows experimenting on a wide range of scenari
os beyond the man-in-the-middle attack tested here, including the case of physic
al attacks.
****
GaitGCI: Generative Counterfactual Intervention for Gait Recognition
Huanzhang Dou, Pengyi Zhang, Wei Su, Yunlong Yu, Yining Lin, Xi Li; Proceedings
of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20
23, pp. 5578-5588
Gait is one of the most promising biometrics that aims to identify pedestrians f
rom their walking patterns. However, prevailing methods are susceptible to confo
unders, resulting in the networks hardly focusing on the regions that reflect ef
fective walking patterns. To address this fundamental problem in gait recognitio
n, we propose a Generative Counterfactual Intervention framework, dubbed GaitGCI
, consisting of Counterfactual Intervention Learning (CIL) and Diversity-Constra
ined Dynamic Convolution (DCDC). CIL leverages causal inference to alleviate the
 impact of confounders by maximizing the likelihood difference between factual/c
ounterfactual attention. DCDC adaptively generates sample-wise factual/counterfa
ctual attention to perceive the sample properties. With matrix decomposition and
 diversity constraint, DCDC guarantees the model's efficiency and effectiveness.
 Extensive experiments indicate that proposed GaitGCI: 1) could effectively focu
s on the discriminative and interpretable regions that reflect gait patterns; 2)
 is model-agnostic and could be plugged into existing models to improve performa
nce with nearly no extra cost; 3) efficiently achieves state-of-the-art performa
nce on arbitrary scenarios (in-the-lab and in-the-wild).
****
Constructing Deep Spiking Neural Networks From Artificial Neural Networks With K
nowledge Distillation
Qi Xu, Yaxin Li, Jiangrong Shen, Jian K. Liu, Huajin Tang, Gang Pan; Proceedings
 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 7886-7895
Spiking neural networks (SNNs) are well known as the brain-inspired models with
high computing efficiency, due to a key component that they utilize spikes as in

formation units, close to the biological neural systems. Although spiking based models are energy efficient by taking advantage of discrete spike signals, their performance is limited by current network structures and their training methods. As discrete signals, typical SNNs cannot apply the gradient descent rules directly into parameters adjustment as artificial neural networks (ANNs). Aiming at this limitation, here we propose a novel method of constructing deep SNN models with knowledge distillation (KD) that uses ANN as teacher model and SNN as student model. Through ANN-SNN joint training algorithm, the student SNN model can learn rich feature information from the teacher ANN model through the KD method, yet it avoids training SNN from scratch when communicating with non-differentiable spikes. Our method can not only build a more efficient deep spiking structure feasibly and reasonably, but use few time steps to train whole model compared to direct training or ANN to SNN methods. More importantly, it has a superb ability of noise immunity for various types of artificial noises and natural signals. The proposed novel method provides efficient ways to improve the performance of SNN through constructing deeper structures in a high-throughput fashion, with potential usage for light and efficient brain-inspired computing of practical scenarios.
****
Understanding and Improving Visual Prompting: A Label-Mapping Perspective
Aochuan Chen, Yuguang Yao, Pin-Yu Chen, Yihua Zhang, Sijia Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19133-19143
We revisit and advance visual prompting (VP), an input prompting technique for vision tasks. VP can reprogram a fixed, pre-trained source model to accomplish downstream tasks in the target domain by simply incorporating universal prompts (in terms of input perturbation patterns) into downstream data points. Yet, it remains elusive why VP stays effective even given a ruleless label mapping (LM) between the source classes and the target classes. Inspired by the above, we ask: How is LM interrelated with VP? And how to exploit such a relationship to improve its accuracy on target tasks? We peer into the influence of LM on VP and provide an affirmative answer that a better 'quality' of LM (assessed by mapping precision and explanation) can consistently improve the effectiveness of VP. This is in contrast to the prior art where the factor of LM was missing. To optimize LM, we propose a new VP framework, termed ILM-VP (iterative label mapping-based visual prompting), which automatically re-maps the source labels to the target labels and progressively improves the target task accuracy of VP. Further, when using a contrastive language-image pretrained (CLIP) model, we propose to integrate an LM process to assist the text prompt selection of CLIP and to improve the target task accuracy. Extensive experiments demonstrate that our proposal significantly outperforms state-of-the-art VP methods. As highlighted below, we show that when reprogramming an ImageNet-pretrained ResNet-18 to 13 target tasks, our method outperforms baselines by a substantial margin, e.g., 7.9% and 6.7% accuracy improvements in transfer learning to the target Flowers102 and CIFAR100 datasets. Besides, our proposal on CLIP-based VP provides 13.7% and 7.1% accuracy improvements on Flowers102 and DTD respectively.
****
Directional Connectivity-Based Segmentation of Medical Images
Ziyun Yang, Sina Farsiu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11525-11535
Anatomical consistency in biomarker segmentation is crucial for many medical image analysis tasks. A promising paradigm for achieving anatomically consistent segmentation via deep networks is incorporating pixel connectivity, a basic concept in digital topology, to model inter-pixel relationships. However, previous works on connectivity modeling have ignored the rich channel-wise directional information in the latent space. In this work, we demonstrate that effective disentanglement of directional sub-space from the shared latent space can significantly enhance the feature representation in the connectivity-based network. To this end, we propose a directional connectivity modeling scheme for segmentation that decouples, tracks, and utilizes the directional information across the network. E

xperiments on various public medical image segmentation benchmarks show the effectiveness of our model as compared to the state-of-the-art methods. Code is available at https://github.com/Zyun-Y/DconnNet.
****
Towards Flexible Multi-Modal Document Models
Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, Kota Yamaguchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14287-14296
Creative workflows for generating graphical documents involve complex inter-related tasks, such as aligning elements, choosing appropriate fonts, or employing aesthetically harmonious colors. In this work, we attempt at building a holistic model that can jointly solve many different design tasks. Our model, which we denote by FlexDM, treats vector graphic documents as a set of multi-modal elements, and learns to predict masked fields such as element type, position, styling attributes, image, or text, using a unified architecture. Through the use of explicit multi-task learning and in-domain pre-training, our model can better capture the multi-modal relationships among the different document fields. Experimental results corroborate that our single FlexDM is able to successfully solve a multitude of different design tasks, while achieving performance that is competitive with task-specific and costly baselines.
****
DegAE: A New Pretraining Paradigm for Low-Level Vision
Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Yu Qiao, Chao Dong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23292-23303
Self-supervised pretraining has achieved remarkable success in high-level vision, but its application in low-level vision remains ambiguous and not well-established. What is the primitive intention of pretraining? What is the core problem of pretraining in low-level vision? In this paper, we aim to answer these essential questions and establish a new pretraining scheme for low-level vision. Specifically, we examine previous pretraining methods in both high-level and low-level vision, and categorize current low-level vision tasks into two groups based on the difficulty of data acquisition: low-cost and high-cost tasks. Existing literature has mainly focused on pretraining for low-cost tasks, where the observed performance improvement is often limited. However, we argue that pretraining is more significant for high-cost tasks, where data acquisition is more challenging. To learn a general low-level vision representation that can improve the performance of various tasks, we propose a new pretraining paradigm called degradation autoencoder (DegAE). DegAE follows the philosophy of designing pretext task for self-supervised pretraining and is elaborately tailored to low-level vision. With DegAE pretraining, SwinIR achieves a 6.88dB performance gain on image dehaze task, while Uformer obtains 3.22dB and 0.54dB improvement on dehaze and derain tasks, respectively.
****
The Differentiable Lens: Compound Lens Search Over Glass Surfaces and Materials for Object Detection
Geoffroi Côté, Fahim Mannan, Simon Thibault, Jean-François Lalonde, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20803-20812
Most camera lens systems are designed in isolation, separately from downstream computer vision methods. Recently, joint optimization approaches that design lenses alongside other components of the image acquisition and processing pipeline--notably, downstream neural networks--have achieved improved imaging quality or better performance on vision tasks. However, these existing methods optimize only a subset of lens parameters and cannot optimize glass materials given their categorical nature. In this work, we develop a differentiable spherical lens simulation model that accurately captures geometrical aberrations. We propose an optimization strategy to address the challenges of lens design--notorious for non-convex loss function landscapes and many manufacturing constraints--that are exacerbated in joint optimization tasks. Specifically, we introduce quantized continuo

us glass variables to facilitate the optimization and selection of glass materials in an end-to-end design context, and couple this with carefully designed constraints to support manufacturability. In automotive object detection, we report improved detection performance over existing designs even when simplifying designs to two- or three-element lenses, despite significantly degrading the image quality.
****
Adversarially Masking Synthetic To Mimic Real: Adaptive Noise Injection for Point Cloud Segmentation Adaptation
Guangrui Li, Guoliang Kang, Xiaohan Wang, Yunchao Wei, Yi Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20464-20474
This paper considers the synthetic-to-real adaptation of point cloud semantic segmentation, which aims to segment the real-world point clouds with only synthetic labels available. Contrary to synthetic data which is integral and clean, point clouds collected by real-world sensors typically contain unexpected and irregular noise because the sensors may be impacted by various environmental conditions. Consequently, the model trained on ideal synthetic data may fail to achieve satisfactory segmentation results on real data. Influenced by such noise, previous adversarial training methods, which are conventional for 2D adaptation tasks, become less effective. In this paper, we aim to mitigate the domain gap caused by target noise via learning to mask the source points during the adaptation procedure. To this end, we design a novel learnable masking module, which takes source features and 3D coordinates as inputs. We incorporate Gumbel-Softmax operation into the masking module so that it can generate binary masks and be trained end-to-end via gradient back-propagation. With the help of adversarial training, the masking module can learn to generate source masks to mimic the pattern of irregular target noise, thereby narrowing the domain gap. We name our method "Adversarial Masking" as adversarial training and learnable masking module depend on each other and cooperate with each other to mitigate the domain gap. Experiments on two synthetic-to-real adaptation benchmarks verify the effectiveness of the proposed method.
****
KERM: Knowledge Enhanced Reasoning for Vision-and-Language Navigation
Xiangyang Li, Zihan Wang, Jiahao Yang, Yaowei Wang, Shuqiang Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2583-2592
Vision-and-language navigation (VLN) is the task to enable an embodied agent to navigate to a remote location following the natural language instruction in real scenes. Most of the previous approaches utilize the entire features or object-centric features to represent navigable candidates. However, these representations are not efficient enough for an agent to perform actions to arrive the target location. As knowledge provides crucial information which is complementary to visible content, in this paper, we propose a Knowledge Enhanced Reasoning Model (KERM) to leverage knowledge to improve agent navigation ability. Specifically, we first retrieve facts (i.e., knowledge described by language descriptions) for the navigation views based on local regions from the constructed knowledge base. The retrieved facts range from properties of a single object (e.g., color, shape) to relationships between objects (e.g., action, spatial position), providing crucial information for VLN. We further present the KERM which contains the purification, fact-aware interaction, and instruction-guided aggregation modules to integrate visual, history, instruction, and fact features. The proposed KERM can automatically select and gather crucial and relevant cues, obtaining more accurate action prediction. Experimental results on the REVERIE, R2R, and SOON datasets demonstrate the effectiveness of the proposed method. The source code is available at https://github.com/XiangyangLi20/KERM.
****
LiDAR-in-the-Loop Hyperparameter Optimization
Félix Goudreault, Dominik Scheuble, Mario Bijelic, Nicolas Robidoux, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit

ion (CVPR), 2023, pp. 13404-13414
LiDAR has become a cornerstone sensing modality for 3D vision. LiDAR systems emit pulses of light into the scene, take measurements of the returned signal, and rely on hardware digital signal processing (DSP) pipelines to construct 3D point clouds from these measurements. The resulting point clouds output by these DSPs are input to downstream 3D vision models -- both, in the form of training datasets or as input at inference time. Existing LiDAR DSPs are composed of cascades of parameterized operations; modifying configuration parameters results in significant changes in the point clouds and consequently the output of downstream methods. Existing methods treat LiDAR systems as fixed black boxes and construct downstream task networks more robust with respect to measurement fluctuations. Departing from this approach, the proposed method directly optimizes LiDAR sensing and DSP parameters for downstream tasks. To investigate the optimization of LiDAR system parameters, we devise a realistic LiDAR simulation method that generates raw waveforms as input to a LiDAR DSP pipeline. We optimize LiDAR parameters for both 3D object detection IoU losses and depth error metrics by solving a nonlinear multi-objective optimization problem with a 0th-order stochastic algorithm. For automotive 3D object detection models, the proposed method outperforms manual expert tuning by 39.5% mean Average Precision (mAP).
****
Local 3D Editing via 3D Distillation of CLIP Knowledge
Junha Hyung, Sungwon Hwang, Daejin Kim, Hyunji Lee, Jaegul Choo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12674-12684
3D content manipulation is an important computer vision task with many real-world applications (e.g., product design, cartoon generation, and 3D Avatar editing). Recently proposed 3D GANs can generate diverse photo-realistic 3D-aware contents using Neural Radiance fields (NeRF). However, manipulation of NeRF still remains a challenging problem since the visual quality tends to degrade after manipulation and suboptimal control handles such as semantic maps are used for manipulations. While text-guided manipulations have shown potential in 3D editing, such approaches often lack locality. To overcome the problems, we propose Local Editing NeRF (LENeRF), which only requires text inputs for fine-grained and localized manipulation. Specifically, we present three add-on modules of LENeRF, the Latent Residual Mapper, the Attention Field Network, and the Deformation Network, which are jointly used for local manipulations of 3D features by estimating a 3D attention field. The 3D attention field is learned in an unsupervised way, by distilling the CLIP's zero-shot mask generation capability to 3D with multi-view guidance. We conduct diverse experiments and thorough evaluations both quantitatively and qualitatively.
****
Abstract Visual Reasoning: An Algebraic Approach for Solving Raven's Progressive Matrices
Jingyi Xu, Tushar Vaidya, Yufei Wu, Saket Chandra, Zhangsheng Lai, Kai Fong Ernest Chong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6715-6724
We introduce algebraic machine reasoning, a new reasoning framework that is well-suited for abstract reasoning. Effectively, algebraic machine reasoning reduces the difficult process of novel problem-solving to routine algebraic computation. The fundamental algebraic objects of interest are the ideals of some suitably initialized polynomial ring. We shall explain how solving Raven's Progressive Matrices (RPMs) can be realized as computational problems in algebra, which combine various well-known algebraic subroutines that include: Computing the Grobner basis of an ideal, checking for ideal containment, etc. Crucially, the additional algebraic structure satisfied by ideals allows for more operations on ideals beyond set-theoretic operations. Our algebraic machine reasoning framework is not only able to select the correct answer from a given answer set, but also able to generate the correct answer with only the question matrix given. Experiments on the I-RAVEN dataset yield an overall 93.2% accuracy, which significantly outperforms the current state-of-the-art accuracy of 77.0% and exceeds human performan

ce at 84.4% accuracy.
****
3D-Aware Conditional Image Synthesis
Kangle Deng, Gengshan Yang, Deva Ramanan, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4434-4445
We propose pix2pix3D, a 3D-aware conditional generative model for controllable photorealistic image synthesis. Given a 2D label map, such as a segmentation or edge map, our model learns to synthesize a corresponding image from different viewpoints. To enable explicit 3D user control, we extend conditional generative models with neural radiance fields. Given widely-available posed monocular image and label map pairs, our model learns to assign a label to every 3D point in addition to color and density, which enables it to render the image and pixel-aligned label map simultaneously. Finally, we build an interactive system that allows users to edit the label map from different viewpoints and generate outputs accordingly.
****
Understanding Deep Generative Models With Generalized Empirical Likelihoods
Suman Ravuri, Mélanie Rey, Shakir Mohamed, Marc Peter Deisenroth; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24395-24405
Understanding how well a deep generative model captures a distribution of high-dimensional data remains an important open challenge. It is especially difficult for certain model classes, such as Generative Adversarial Networks and Diffusion Models, whose models do not admit exact likelihoods. In this work, we demonstrate that generalized empirical likelihood (GEL) methods offer a family of diagnostic tools that can identify many deficiencies of deep generative models (DGMs). We show, with appropriate specification of moment conditions, that the proposed method can identify which modes have been dropped, the degree to which DGMs are mode imbalanced, and whether DGMs sufficiently capture intra-class diversity. We show how to combine techniques from Maximum Mean Discrepancy and Generalized Empirical Likelihood to create not only distribution tests that retain per-sample interpretability, but also metrics that include label information. We find that such tests predict the degree of mode dropping and mode imbalance up to 60% better than metrics such as improved precision/recall.
****
ABCD: Arbitrary Bitwise Coefficient for De-Quantization
Woo Kyoung Han, Byeonghun Lee, Sang Hyun Park, Kyong Hwan Jin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5876-5885
Modern displays and contents support more than 8bits image and video. However, bit-starving situations such as compression codecs make low bit-depth (LBD) images (<8bits), occurring banding and blurry artifacts. Previous bit depth expansion (BDE) methods still produce unsatisfactory high bit-depth (HBD) images. To this end, we propose an implicit neural function with a bit query to recover de-quantized images from arbitrarily quantized inputs. We develop a phasor estimator to exploit the information of the nearest pixels. Our method shows superior performance against prior BDE methods on natural and animation images. We also demonstrate our model on YouTube UGC datasets for de-banding. Our source code is available at https://github.com/WooKyoungHan/ABCD
****
Event-Based Blurry Frame Interpolation Under Blind Exposure
Wenming Weng, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1588-1598
Restoring sharp high frame-rate videos from low frame-rate blurry videos is a challenging problem. Existing blurry frame interpolation methods assume a predefined and known exposure time, which suffer from severe performance drop when applied to videos captured in the wild. In this paper, we study the problem of blurry frame interpolation under blind exposure with the assistance of an event camera. The high temporal resolution of the event camera is beneficial to obtain the e

xposure prior that is lost during the imaging process. Besides, sharp frames can be restored using event streams and blurry frames relying on the mutual constraint among them. Therefore, we first propose an exposure estimation strategy guided by event streams to estimate the lost exposure prior, transforming the blind exposure problem well-posed. Second, we propose to model the mutual constraint with a temporal-exposure control strategy through iterative residual learning. Our blurry frame interpolation method achieves a distinct performance boost over existing methods on both synthetic and self-collected real-world datasets under blind exposure.
****

Human Body Shape Completion With Implicit Shape and Flow Learning

Boyao Zhou, Di Meng, Jean-Sébastien Franco, Edmond Boyer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12901-12911

In this paper, we investigate how to complete human body shape models by combining shape and flow estimation given two consecutive depth images. Shape completion is a challenging task in computer vision that is highly under-constrained when considering partial depth observations. Besides model based strategies that exploit strong priors, and consequently struggle to preserve fine geometric details, learning based approaches build on weaker assumptions and can benefit from efficient implicit representations. We adopt such a representation and explore how the motion flow between two consecutive frames can contribute to the shape completion task. In order to effectively exploit the flow information, our architecture combines both estimations and implements two features for robustness: First, an all-to-all attention module that encodes the correlation between points in the same frame and between corresponding points in different frames; Second, a coarse-dense to fine-sparse strategy that balances the representation ability and the computational cost. Our experiments demonstrate that the flow actually benefits human body model completion. They also show that our method outperforms the state-of-the-art approaches for shape completion on 2 benchmarks, considering different human shapes, poses, and clothing.
****

Spider GAN: Leveraging Friendly Neighbors To Accelerate GAN Training

Siddarth Asokan, Chandra Sekhar Seelamantula; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3883-3893

Training Generative adversarial networks (GANs) stably is a challenging task. The generator in GANs transform noise vectors, typically Gaussian distributed, into realistic data such as images. In this paper, we propose a novel approach for training GANs with images as inputs, but without enforcing any pairwise constraints. The intuition is that images are more structured than noise, which the generator can leverage to learn a more robust transformation. The process can be made efficient by identifying closely related datasets, or a "friendly neighborhood" of the target distribution, inspiring the moniker, Spider GAN. To define friendly neighborhoods leveraging proximity between datasets, we propose a new measure called the signed inception distance (SID), inspired by the polyharmonic kernel. We show that the Spider GAN formulation results in faster convergence, as the generator can discover correspondence even between seemingly unrelated datasets, for instance, between Tiny-ImageNet and CelebA faces. Further, we demonstrate cascading Spider GAN, where the output distribution from a pre-trained GAN generator is used as the input to the subsequent network. Effectively, transporting one distribution to another in a cascaded fashion until the target is learnt -- a new flavor of transfer learning. We demonstrate the efficacy of the Spider approach on DCGAN, conditional GAN, PGGAN, StyleGAN2 and StyleGAN3. The proposed approach achieves state-of-the-art Frechet inception distance (FID) values, with one-fifth of the training iterations, in comparison to their baseline counterparts on high-resolution small datasets such as MetFaces, Ukiyo-E Faces and AFHQ-Cats.
****

CLIPPING: Distilling CLIP-Based Models With a Student Base for Video-Language Retrieval

Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei L u, Youliang Yan; Proceedings of the IEEE/CVF Conference on Computer Vision and P attern Recognition (CVPR), 2023, pp. 18983-18992
Pre-training a vison-language model and then fine-tuning it on downstream tasks have become a popular paradigm. However, pre-trained vison-language models with the Transformer architecture usually take long inference time. Knowledge distill ation has been an efficient technique to transfer the capability of a large mode l to a small one while maintaining the accuracy, which has achieved remarkable s uccess in natural language processing. However, it faces many problems when appl ying KD to the multi-modality applications. In this paper, we propose a novel kn owledge distillation method, named CLIPPING, where the plentiful knowledge of a large teacher model that has been fine-tuned for video-language tasks with the p owerful pre-trained CLIP can be effectively transferred to a small student only at the fine-tuning stage. Especially, a new layer-wise alignment with the studen t as the base is proposed for knowledge distillation of the intermediate layers in CLIPPING, which enables the student's layers to be the bases of the teacher, and thus allows the student to fully absorb the knowledge of the teacher. CLIPPI NG with MobileViT-v2 as the vison encoder without any vison-language pre-trainin g achieves 88.1%-95.3% of the performance of its teacher on three video-language retrieval benchmarks, with its vison encoder being 19.5x smaller. CLIPPING also significantly outperforms a state-of-the-art small baseline (ALL-in-one-B) on t he MSR-VTT dataset, obtaining relatively 7.4% performance gain, with 29% fewer p arameters and 86.9% fewer flops. Moreover, CLIPPING is comparable or even superi or to many large pre-training models.
****
ScaleDet: A Scalable Multi-Dataset Object Detector
Yanbei Chen, Manchen Wang, Abhay Mittal, Zhenlin Xu, Paolo Favaro, Joseph Tighe, Davide Modolo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pa ttern Recognition (CVPR), 2023, pp. 7288-7297
Multi-dataset training provides a viable solution for exploiting heterogeneous l arge-scale datasets without extra annotation cost. In this work, we propose a sc alable multi-dataset detector (ScaleDet) that can scale up its generalization ac ross datasets when increasing the number of training datasets. Unlike existing m ulti-dataset learners that mostly rely on manual relabelling efforts or sophisti cated optimizations to unify labels across datasets, we introduce a simple yet s calable formulation to derive a unified semantic label space for multi-dataset t raining. ScaleDet is trained by visual-textual alignment to learn the label assi gnment with label semantic similarities across datasets. Once trained, ScaleDet can generalize well on any given upstream and downstream datasets with seen and unseen classes. We conduct extensive experiments using LVIS, COCO, Objects365, O penImages as upstream datasets, and 13 datasets from Object Detection in the Wil d (ODinW) as downstream datasets. Our results show that ScaleDet achieves compel ling strong model performance with an mAP of 50.7 on LVIS, 58.8 on COCO, 46.8 on Objects365, 76.2 on OpenImages, and 71.8 on ODinW, surpassing state-of-the-art detectors with the same backbone.
****
Unbiased Multiple Instance Learning for Weakly Supervised Video Anomaly Detectio n
Hui Lv, Zhongqi Yue, Qianru Sun, Bin Luo, Zhen Cui, Hanwang Zhang; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 3, pp. 8022-8031
Weakly Supervised Video Anomaly Detection (WSVAD) is challenging because the bin ary anomaly label is only given on the video level, but the output requires snip pet-level predictions. So, Multiple Instance Learning (MIL) is prevailing in WSV AD. However, MIL is notoriously known to suffer from many false alarms because t he snippet-level detector is easily biased towards the abnormal snippets with si mple context, confused by the normality with the same bias, and missing the anom aly with a different pattern. To this end, we propose a new MIL framework: Unbia sed MIL (UMIL), to learn unbiased anomaly features that improve WSVAD. At each M IL training iteration, we use the current detector to divide the samples into tw

o groups with different context biases: the most confident abnormal/normal snippets and the rest ambiguous ones. Then, by seeking the invariant features across the two sample groups, we can remove the variant context biases. Extensive experiments on benchmarks UCF-Crime and TAD demonstrate the effectiveness of our UMIL. Our code is provided at https://github.com/ktr-hubrt/UMIL.
****
BEVHeight: A Robust Framework for Vision-Based Roadside 3D Object Detection
Lei Yang, Kaicheng Yu, Tao Tang, Jun Li, Kun Yuan, Li Wang, Xinyu Zhang, Peng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21611-21620
While most recent autonomous driving system focuses on developing perception methods on ego-vehicle sensors, people tend to overlook an alternative approach to leverage intelligent roadside cameras to extend the perception ability beyond the visual range. We discover that the state-of-the-art vision-centric bird's eye view detection methods have inferior performances on roadside cameras. This is because these methods mainly focus on recovering the depth regarding the camera center, where the depth difference between the car and the ground quickly shrinks while the distance increases. In this paper, we propose a simple yet effective approach, dubbed BEVHeight, to address this issue. In essence, instead of predicting the pixel-wise depth, we regress the height to the ground to achieve a distance-agnostic formulation to ease the optimization process of camera-only perception methods. On popular 3D detection benchmarks of roadside cameras, our method surpasses all previous vision-centric methods by a significant margin. The code is available at https://github.com/ADLab-AutoDrive/BEVHeight.
****
Towards Unbiased Volume Rendering of Neural Implicit Surfaces With Geometry Priors
Yongqiang Zhang, Zhipeng Hu, Haoqian Wu, Minda Zhao, Lincheng Li, Zhengxia Zou, Changjie Fan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4359-4368
Learning surface by neural implicit rendering has been a promising way for multi-view reconstruction in recent years. Existing neural surface reconstruction methods, such as NeuS and VolSDF, can produce reliable meshes from multi-view posed images. Although they build a bridge between volume rendering and Signed Distance Function (SDF), the accuracy is still limited. In this paper, we argue that this limited accuracy is due to the bias of their volume rendering strategies, especially when the viewing direction is close to be tangent to the surface. We revise and provide an additional condition for the unbiased volume rendering. Following this analysis, we propose a new rendering method by scaling the SDF field with the angle between the viewing direction and the surface normal vector. Experiments on simulated data indicate that our rendering method reduces the bias of SDF-based volume rendering. Moreover, there still exists non-negligible bias when the learnable standard deviation of SDF is large at early stage, which means that it is hard to supervise the rendered depth with depth priors. Alternatively we supervise zero-level set with surface points obtained from a pre-trained Multi-View Stereo network. We evaluate our method on the DTU dataset and show that it outperforms the state-of-the-arts neural implicit surface methods without mask supervision.
****
Modular Memorability: Tiered Representations for Video Memorability Prediction
Théo Dumont, Juan Segundo Hevia, Camilo L. Fosco; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10751-10760
The question of how to best estimate the memorability of visual content is currently a source of debate in the memorability community. In this paper, we propose to explore how different key properties of images and videos affect their consolidation into memory. We analyze the impact of several features and develop a model that emulates the most important parts of a proposed "pathway to memory": a simple but effective way of representing the different hurdles that new visual content needs to surpass to stay in memory. This framework leads to the construct

ion of our M3-S model, a novel memorability network that processes input videos in a modular fashion. Each module of the network emulates one of the four key steps of the pathway to memory: raw encoding, scene understanding, event understanding and memory consolidation. We find that the different representations learned by our modules are non-trivial and substantially different from each other. Additionally, we observe that certain representations tend to perform better at the task of memorability prediction than others, and we introduce an in-depth ablation study to support our results. Our proposed approach surpasses the state of the art on the two largest video memorability datasets and opens the door to new applications in the field.
****
Weakly-Supervised Domain Adaptive Semantic Segmentation With Prototypical Contrastive Learning
Anurag Das, Yongqin Xian, Dengxin Dai, Bernt Schiele; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15434-15443
There has been a lot of effort in improving the performance of unsupervised domain adaptation for semantic segmentation task, however there is still a huge gap in performance when compared with supervised learning. In this work, we propose a common framework to use different weak labels, e.g. image, point and coarse labels from target domain to reduce this performance gap. Specifically, we propose to learn better prototypes that are representative class features, by exploiting these weak labels. We use these improved prototypes for contrastive alignment of class features. In particular, we perform two different feature alignments, first, we align pixel features with prototypes within each domain and second, we align pixel features from source to prototype of target domain in an asymmetric way. This asymmetric alignment is beneficial as it preserves the target features during training, which is essential when weak labels are available from target domain. Our experiments on standard benchmarks shows that our framework achieves significant improvement compared to existing works and is able to reduce the performance gap with supervised learning.
****
Language-Guided Music Recommendation for Video via Prompt Analogies
Daniel McKee, Justin Salamon, Josef Sivic, Bryan Russell; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14784-14793
We propose a method to recommend music for an input video while allowing a user to guide music selection with free-form natural language. A key challenge of this problem setting is that existing music video datasets provide the needed (video, music) training pairs, but lack text descriptions of the music. This work addresses this challenge with the following three contributions. First, we propose a text-synthesis approach that relies on an analogy-based prompting procedure to generate natural language music descriptions from a large-scale language model (BLOOM-176B) given pre-trained music tagger outputs and a small number of human text descriptions. Second, we use these synthesized music descriptions to train a new trimodal model, which fuses text and video input representations to query music samples. For training, we introduce a text dropout regularization mechanism which we show is critical to model performance. Our model design allows for the retrieved music audio to agree with the two input modalities by matching visual style depicted in the video and musical genre, mood, or instrumentation described in the natural language query. Third, to evaluate our approach, we collect a testing dataset for our problem by annotating a subset of 4k clips from the YT8M-MusicVideo dataset with natural language music descriptions which we make publicly available. We show that our approach can match or exceed the performance of prior methods on video-to-music retrieval while significantly improving retrieval accuracy when using text guidance.
****
Re2TAL: Rewiring Pretrained Video Backbones for Reversible Temporal Action Localization
Chen Zhao, Shuming Liu, Karttikeya Mangalam, Bernard Ghanem; Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10637-10647

Temporal action localization (TAL) requires long-form reasoning to predict actions of various durations and complex content. Given limited GPU memory, training TAL end to end (i.e., from videos to predictions) on long videos is a significant challenge. Most methods can only train on pre-extracted features without optimizing them for the localization problem, consequently limiting localization performance. In this work, to extend the potential in TAL networks, we propose a novel end-to-end method Re2TAL, which rewires pretrained video backbones for reversible TAL. Re2TAL builds a backbone with reversible modules, where the input can be recovered from the output such that the bulky intermediate activations can be cleared from memory during training. Instead of designing one single type of reversible module, we propose a network rewiring mechanism, to transform any module with a residual connection to a reversible module without changing any parameters. This provides two benefits: (1) a large variety of reversible networks are easily obtained from existing and even future model designs, and (2) the reversible models require much less training effort as they reuse the pre-trained parameters of their original non-reversible versions. Re2TAL, only using the RGB modality, reaches 37.01% average mAP on ActivityNet-v1.3, a new state-of-the-art record, and mAP 64.9% at tIoU=0.5 on THUMOS-14, outperforming all other RGB-only methods. Code is available at https://github.com/coolbay/Re2TAL.

****

Neuro-Modulated Hebbian Learning for Fully Test-Time Adaptation

Yushun Tang, Ce Zhang, Heng Xu, Shuoshuo Chen, Jie Cheng, Luziwei Leng, Qinghai Guo, Zhihai He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3728-3738

Fully test-time adaptation aims to adapt the network model based on sequential analysis of input samples during the inference stage to address the cross-domain performance degradation problem of deep neural networks. We take inspiration from the biological plausibility learning where the neuron responses are tuned based on a local synapse-change procedure and activated by competitive lateral inhibition rules. Based on these feed-forward learning rules, we design a soft Hebbian learning process which provides an unsupervised and effective mechanism for online adaptation. We observe that the performance of this feed-forward Hebbian learning for fully test-time adaptation can be significantly improved by incorporating a feedback neuro-modulation layer. It is able to fine-tune the neuron responses based on the external feedback generated by the error back-propagation from the top inference layers. This leads to our proposed neuro-modulated Hebbian learning (NHL) method for fully test-time adaptation. With the unsupervised feed-forward soft Hebbian learning being combined with a learned neuro-modulator to capture feedback from external responses, the source model can be effectively adapted during the testing process. Experimental results on benchmark datasets demonstrate that our proposed method can significantly improve the adaptation performance of network models and outperforms existing state-of-the-art methods.

****

NeRFLight: Fast and Light Neural Radiance Fields Using a Shared Feature Grid

Fernando Rivas-Manzaneque, Jorge Sierra-Acosta, Adrian Penate-Sanchez, Francesc Moreno-Noguer, Angela Ribeiro; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12417-12427

While original Neural Radiance Fields (NeRF) have shown impressive results in modeling the appearance of a scene with compact MLP architectures, they are not able to achieve real-time rendering. This has been recently addressed by either baking the outputs of NeRF into a data structure or arranging trainable parameters in an explicit feature grid. These strategies, however, significantly increase the memory footprint of the model which prevents their deployment on bandwidth-constrained applications. In this paper, we extend the grid-based approach to achieve real-time view synthesis at more than 150 FPS using a lightweight model. Our main contribution is a novel architecture in which the density field of NeRF-based representations is split into N regions and the density is modeled using N different decoders which reuse the same feature grid. This results in a smaller

grid where each feature is located in more than one spatial position, forcing th em to learn a compact representation that is valid for different parts of the sc ene. We further reduce the size of the final model by disposing of the features symmetrically on each region, which favors feature pruning after training while also allowing smooth gradient transitions between neighboring voxels. An exhaust ive evaluation demonstrates that our method achieves real-time performance and q uality metrics on a pair with state-of-the-art with an improvement of more than 2x in the FPS/MB ratio.
****
MVImgNet: A Large-Scale Dataset of Multi-View Images
Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizh eng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, Guanying Chen, Shuguang C ui, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9150-9161
Being data-driven is one of the most iconic properties of deep learning algorith ms. The birth of ImageNet drives a remarkable trend of "learning from large-scal e data" in computer vision. Pretraining on ImageNet to obtain rich universal rep resentations has been manifested to benefit various 2D visual tasks, and becomes a standard in 2D vision. However, due to the laborious collection of real-world 3D data, there is yet no generic dataset serving as a counterpart of ImageNet i n 3D vision, thus how such a dataset can impact the 3D community is unraveled. T o remedy this defect, we introduce MVImgNet, a large-scale dataset of multi-view images, which is highly convenient to gain by shooting videos of real-world obj ects in human daily life. It contains 6.5 million frames from 219,188 videos cro ssing objects from 238 classes, with rich annotations of object masks, camera pa rameters, and point clouds. The multi-view attribute endows our dataset with 3D-aware signals, making it a soft bridge between 2D and 3D vision. We conduct pilo t studies for probing the potential of MVImgNet on a variety of 3D and 2D visual tasks, including radiance field reconstruction, multi-view stereo, and view-con sistent image understanding, where MVImgNet demonstrates promising performance, remaining lots of possibilities for future explorations. Besides, via dense reco nstruction on MVImgNet, a 3D object point cloud dataset is derived, called MVPNe t, covering 87,200 samples from 150 categories, with the class label on each poi nt cloud. Experiments show that MVPNet can benefit the real-world 3D object clas sification while posing new challenges to point cloud understanding. MVImgNet an d MVPNet will be publicly available, hoping to inspire the broader vision commun ity.
****
LASP: Text-to-Text Optimization for Language-Aware Soft Prompting of Vision & La nguage Models
Adrian Bulat, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23232-23241
Soft prompt learning has recently emerged as one of the methods of choice for ad apting V&L models to a downstream task using a few training examples. However, c urrent methods significantly overfit the training data, suffering from large acc uracy degradation when tested on unseen classes from the same domain. To this en d, in this paper, we make the following 4 contributions: (1) To alleviate base c lass overfitting, we propose a novel Language-Aware Soft Prompting (LASP) learni ng method by means of a text-to-text cross-entropy loss that maximizes the proba bility of the learned prompts to be correctly classified with respect to pre-def ined hand-crafted textual prompts. (2) To increase the representation capacity o f the prompts, we propose grouped LASP where each group of prompts is optimized with respect to a separate subset of textual prompts. (3) We identify a visual-l anguage misalignment introduced by prompt learning and LASP, and more importantl y, propose a re-calibration mechanism to address it. (4) We show that LASP is in herently amenable to including, during training, virtual classes, i.e. class nam es for which no visual samples are available, further increasing the robustness of the learned prompts. Through evaluations on 11 datasets, we show that our app roach (a) significantly outperforms all prior works on soft prompting, and (b) m atches and surpasses, for the first time, the accuracy on novel classes obtained

by hand-crafted prompts and CLIP for 8 out of 11 test datasets. Code will be made available.
****

Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization
Shichao Dong, Jin Wang, Renhe Ji, Jiajun Liang, Haoqiang Fan, Zheng Ge; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3994-4004
In this paper, we analyse the generalization ability of binary classifiers for the task of deepfake detection. We find that the stumbling block to their generalization is caused by the unexpected learned identity representation on images. Termed as the Implicit Identity Leakage, this phenomenon has been qualitatively and quantitatively verified among various DNNs. Furthermore, based on such understanding, we propose a simple yet effective method named the ID-unaware Deepfake Detection Model to reduce the influence of this phenomenon. Extensive experimental results demonstrate that our method outperforms the state-of-the-art in both in-dataset and cross-dataset evaluation. The code is available at https://github.com/megvii-research/CADDM.
****

Learning Federated Visual Prompt in Null Space for MRI Reconstruction
Chun-Mei Feng, Bangjun Li, Xinxing Xu, Yong Liu, Huazhu Fu, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8064-8073
Federated Magnetic Resonance Imaging (MRI) reconstruction enables multiple hospitals to collaborate distributedly without aggregating local data, thereby protecting patient privacy. However, the data heterogeneity caused by different MRI protocols, insufficient local training data, and limited communication bandwidth inevitably impair global model convergence and updating. In this paper, we propose a new algorithm, FedPR, to learn federated visual prompts in the null space of global prompt for MRI reconstruction. FedPR is a new federated paradigm that adopts a powerful pre-trained model while only learning and communicating the prompts with few learnable parameters, thereby significantly reducing communication costs and achieving competitive performance on limited local data. Moreover, to deal with catastrophic forgetting caused by data heterogeneity, FedPR also updates efficient federated visual prompts that project the local prompts into an approximate null space of the global prompt, thereby suppressing the interference of gradients on the server performance. Extensive experiments on federated MRI show that FedPR significantly outperforms state-of-the-art FL algorithms with < 6% of communication costs when given the limited amount of local data.
****

A New Benchmark: On the Utility of Synthetic Data With Blender for Bare Supervised Learning and Downstream Domain Adaptation
Hui Tang, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15954-15964
Deep learning in computer vision has achieved great success with the price of large-scale labeled training data. However, exhaustive data annotation is impracticable for each task of all domains of interest, due to high labor costs and unguaranteed labeling accuracy. Besides, the uncontrollable data collection process produces non-IID training and test data, where undesired duplication may exist. All these nuisances may hinder the verification of typical theories and exposure to new findings. To circumvent them, an alternative is to generate synthetic data via 3D rendering with domain randomization. We in this work push forward along this line by doing profound and extensive research on bare supervised learning and downstream domain adaptation. Specifically, under the well-controlled, IID data setting enabled by 3D rendering, we systematically verify the typical, important learning insights, e.g., shortcut learning, and discover the new laws of various data regimes and network architectures in generalization. We further investigate the effect of image formation factors on generalization, e.g., object scale, material texture, illumination, camera viewpoint, and background in a 3D scene. Moreover, we use the simulation-to-reality adaptation as a downstream task

for comparing the transferability between synthetic and real data when used for pre-training, which demonstrates that synthetic data pre-training is also promis ing to improve real test results. Lastly, to promote future research, we develop a new large-scale synthetic-to-real benchmark for image classification, termed S2RDA, which provides more significant challenges for transfer from simulation t o reality. The code and datasets are available at https://github.com/huitangtang /On_the_Utility_of_Synthetic_Data.
****
Data-Driven Feature Tracking for Event Cameras
Nico Messikommer, Carter Fang, Mathias Gehrig, Davide Scaramuzza; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 , pp. 5642-5651
Because of their high temporal resolution, increased resilience to motion blur, and very sparse output, event cameras have been shown to be ideal for low-latenc y and low-bandwidth feature tracking, even in challenging scenarios. Existing fe ature tracking methods for event cameras are either handcrafted or derived from first principles but require extensive parameter tuning, are sensitive to noise, and do not generalize to different scenarios due to unmodeled effects. To tackl e these deficiencies, we introduce the first data-driven feature tracker for eve nt cameras, which leverages low-latency events to track features detected in a g rayscale frame. We achieve robust performance via a novel frame attention module , which shares information across feature tracks. By directly transferring zero-shot from synthetic to real data, our data-driven tracker outperforms existing a pproaches in relative feature age by up to 120% while also achieving the lowest latency. This performance gap is further increased to 130% by adapting our track er to real data with a novel self-supervision strategy.
****
Temporal Consistent 3D LiDAR Representation Learning for Semantic Perception in Autonomous Driving
Lucas Nunes, Louis Wiesmann, Rodrigo Marcuzzi, Xieyuanli Chen, Jens Behley, Cyri ll Stachniss; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt ern Recognition (CVPR), 2023, pp. 5217-5228
Semantic perception is a core building block in autonomous driving, since it pro vides information about the drivable space and location of other traffic partici pants. For learning-based perception, often a large amount of diverse training d ata is necessary to achieve high performance. Data labeling is usually a bottlen eck for developing such methods, especially for dense prediction tasks, e.g., se mantic segmentation or panoptic segmentation. For 3D LiDAR data, the annotation process demands even more effort than for images. Especially in autonomous drivi ng, point clouds are sparse, and objects appearance depends on its distance from the sensor, making it harder to acquire large amounts of labeled training data. This paper aims at taking an alternative path proposing a self-supervised repre sentation learning method for 3D LiDAR data. Our approach exploits the vehicle m otion to match objects across time viewed in different scans. We then train a mo del to maximize the point-wise feature similarities from points of the associate d object in different scans, which enables to learn a consistent representation across time. The experimental results show that our approach performs better tha n previous state-of-the-art self-supervised representation learning methods when fine-tuning to different downstream tasks. We furthermore show that with only 1 0% of labeled data, a network pre-trained with our approach can achieve better p erformance than the same network trained from scratch with all labels for semant ic segmentation on SemanticKITTI.
****
AutoAD: Movie Description in Context
Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, Andrew Zisserman; Pro ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18930-18940
The objective of this paper is an automatic Audio Description (AD) model that in gests movies and outputs AD in text form. Generating high-quality movie AD is ch allenging due to the dependency of the descriptions on context, and the limited

amount of training data available. In this work, we leverage the power of pretra
ined foundation models, such as GPT and CLIP, and only train a mapping network t
hat bridges the two models for visually-conditioned text generation. In order to
 obtain high-quality AD, we make the following four contributions: (i) we incorp
orate context from the movie clip, AD from previous clips, as well as the subtit
les; (ii) we address the lack of training data by pretraining on large-scale dat
asets, where visual or contextual information is unavailable, e.g. text-only AD
without movies or visual captioning datasets without context; (iii) we improve o
n the currently available AD datasets, by removing label noise in the MAD datase
t, and adding character naming information; and (iv) we obtain strong results on
 the movie AD task compared with previous methods.
****
DiffTalk: Crafting Diffusion Models for Generalized Audio-Driven Portraits Anima
tion
Shuai Shen, Wenliang Zhao, Zibin Meng, Wanhua Li, Zheng Zhu, Jie Zhou, Jiwen Lu;
 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti
on (CVPR), 2023, pp. 1982-1991
Talking head synthesis is a promising approach for the video production industry
. Recently, a lot of effort has been devoted in this research area to improve th
e generation quality or enhance the model generalization. However, there are few
 works able to address both issues simultaneously, which is essential for practi
cal applications. To this end, in this paper, we turn attention to the emerging
powerful Latent Diffusion Models, and model the Talking head generation as an au
dio-driven temporally coherent denoising process (DiffTalk). More specifically,
instead of employing audio signals as the single driving factor, we investigate
the control mechanism of the talking face, and incorporate reference face images
 and landmarks as conditions for personality-aware generalized synthesis. In thi
s way, the proposed DiffTalk is capable of producing high-quality talking head v
ideos in synchronization with the source audio, and more importantly, it can be
naturally generalized across different identities without any further fine-tunin
g. Additionally, our DiffTalk can be gracefully tailored for higher-resolution s
ynthesis with negligible extra computational cost. Extensive experiments show th
at the proposed DiffTalk efficiently synthesizes high-fidelity audio-driven talk
ing head videos for generalized novel identities. For more video results, please
 refer to https://sstzal.github.io/DiffTalk/.
****
Autoregressive Visual Tracking
Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, Yihong Gong; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
9697-9706
We present ARTrack, an autoregressive framework for visual object tracking. ARTr
ack tackles tracking as a coordinate sequence interpretation task that estimates
 object trajectories progressively, where the current estimate is induced by pre
vious states and in turn affects subsequences. This time-autoregressive approach
 models the sequential evolution of trajectories to keep tracing the object acro
ss frames, making it superior to existing template matching based trackers that
only consider the per-frame localization accuracy. ARTrack is simple and direct,
 eliminating customized localization heads and post-processings. Despite its sim
plicity, ARTrack achieves state-of-the-art performance on prevailing benchmark d
atasets.
****
SceneComposer: Any-Level Semantic Image Synthesis
Yu Zeng, Zhe Lin, Jianming Zhang, Qing Liu, John Collomosse, Jason Kuen, Vishal
M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2023, pp. 22468-22478
We propose a new framework for conditional image synthesis from semantic layouts
 of any precision levels, ranging from pure text to a 2D semantic canvas with pr
ecise shapes. More specifically, the input layout consists of one or more semant
ic regions with free-form text descriptions and adjustable precision levels, whi
ch can be set based on the desired controllability. The framework naturally redu

ces to text-to-image (T2I) at the lowest level with no shape information, and it becomes segmentation-to-image (S2I) at the highest level. By supporting the levels in-between, our framework is flexible in assisting users of different drawing expertise and at different stages of their creative workflow. We introduce several novel techniques to address the challenges coming with this new setup, including a pipeline for collecting training data; a precision-encoded mask pyramid and a text feature map representation to jointly encode precision level, semantics, and composition information; and a multi-scale guided diffusion model to synthesize images. To evaluate the proposed method, we collect a test dataset containing user-drawn layouts with diverse scenes and styles. Experimental results show that the proposed method can generate high-quality images following the layout at given precision, and compares favorably against existing methods. Project page https://zengxianyu.github.io/scenec/
****
Visual Query Tuning: Towards Effective Usage of Intermediate Representations for Parameter and Memory Efficient Transfer Learning
Cheng-Hao Tu, Zheda Mai, Wei-Lun Chao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7725-7735
Intermediate features of a pre-trained model have been shown informative for making accurate predictions on downstream tasks, even if the model backbone is frozen. The key challenge is how to utilize them, given the gigantic amount. We propose visual query tuning (VQT), a simple yet effective approach to aggregate intermediate features of Vision Transformers. Through introducing a handful of learnable "query" tokens to each layer, VQT leverages the inner workings of Transformers to "summarize" rich intermediate features of each layer, which can then be used to train the prediction heads of downstream tasks. As VQT keeps the intermediate features intact and only learns to combine them, it enjoys memory efficiency in training, compared to many other parameter-efficient fine-tuning approaches that learn to adapt features and need back-propagation through the entire backbone. This also suggests the complementary role between VQT and those approaches in transfer learning. Empirically, VQT consistently surpasses the state-of-the-art approach that utilizes intermediate features for transfer learning and outperforms full fine-tuning in many cases. Compared to parameter-efficient approaches that adapt features, VQT achieves much higher accuracy under memory constraints. Most importantly, VQT is compatible with these approaches to attain higher accuracy, making it a simple add-on to further boost transfer learning.
****
MaPLe: Multi-Modal Prompt Learning
Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19113-19122
Pre-trained vision-language (V-L) models such as CLIP have shown excellent generalization ability to downstream tasks. However, they are sensitive to the choice of input text prompts and require careful selection of prompt templates to perform well. Inspired by the Natural Language Processing (NLP) literature, recent CLIP adaptation approaches learn prompts as the textual inputs to fine-tune CLIP for downstream tasks. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. In this work, we propose Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions. Further, we learn separate prompts across different early stages to progressively model the stage-wise feature relationships to allow rich context learning. We evaluate the effectiveness of our approach on three representative tasks of generalization to novel classes, new target datasets and unseen domain shifts. Compared with the state-of-the-art method Co-CoOp, MaPLe exhibits favorable performance and achieves an absolute gain of 3.45% on novel classes and 2.72% on overall harmonic-mean, averaged over 11 diverse image recognition datasets

. Our code and pre-trained models are available at https://github.com/muzairkhattak/multimodal-prompt-learning.
****
Unsupervised Domain Adaption With Pixel-Level Discriminator for Image-Aware Layout Generation
Chenchen Xu, Min Zhou, Tiezheng Ge, Yuning Jiang, Weiwei Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10114-10123
Layout is essential for graphic design and poster generation. Recently, applying deep learning models to generate layouts has attracted increasing attention. This paper focuses on using the GAN-based model conditioned on image contents to generate advertising poster graphic layouts, which requires an advertising poster layout dataset with paired product images and graphic layouts. However, the paired images and layouts in the existing dataset are collected by inpainting and annotating posters, respectively. There exists a domain gap between inpainted posters (source domain data) and clean product images (target domain data). Therefore, this paper combines unsupervised domain adaption techniques to design a GAN with a novel pixel-level discriminator (PD), called PDA-GAN, to generate graphic layouts according to image contents. The PD is connected to the shallow level feature map and computes the GAN loss for each input-image pixel. Both quantitative and qualitative evaluations demonstrate that PDA-GAN can achieve state-of-the-art performances and generate high-quality image-aware graphic layouts for advertising posters.
****
Compressing Volumetric Radiance Fields to 1 MB
Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, Liefeng Bo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4222-4231
Approximating radiance fields with discretized volumetric grids is one of promising directions for improving NeRFs, represented by methods like DVGO, Plenoxels and TensoRF, which achieve super-fast training convergence and real-time rendering. However, these methods typically require a tremendous storage overhead, costing up to hundreds of megabytes of disk space and runtime memory for a single scene. We address this issue in this paper by introducing a simple yet effective framework, called vector quantized radiance fields (VQRF), for compressing these volume-grid-based radiance fields. We first present a robust and adaptive metric for estimating redundancy in grid models and performing voxel pruning by better exploring intermediate outputs of volumetric rendering. A trainable vector quantization is further proposed to improve the compactness of grid models. In combination with an efficient joint tuning strategy and post-processing, our method can achieve a compression ratio of 100x by reducing the overall model size to 1 MB with negligible loss on visual quality. Extensive experiments demonstrate that the proposed framework is capable of achieving unrivaled performance and well generalization across multiple methods with distinct volumetric structures, facilitating the wide use of volumetric radiance fields methods in real-world applications. Code is available at https://github.com/AlgoHunt/VQRF.
****
Real-Time 6K Image Rescaling With Rate-Distortion Optimization
Chenyang Qi, Xin Yang, Ka Leong Cheng, Ying-Cong Chen, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14092-14101
The task of image rescaling aims at embedding an high-resolution (HR) image into a low-resolution (LR) one that can contain embedded information for HR image reconstruction. Existing image rescaling methods do not optimize the LR image file size and recent flow-based rescaling methods are not real-time yet for HR image reconstruction (e.g., 6K). To address these two challenges, we propose a novel framework (HyperThumbnail) for real-time 6K rate-distortion-aware image rescaling. Our HyperThumbnail first embeds an HR image into a JPEG LR image (thumbnail) by an encoder with our proposed learnable JPEG quantization module, which optimizes the file size of the embedding LR JPEG image. Then, an efficient decoder rec

onstructs a high-fidelity HR (6K) image from the LR one in real time. Extensive experiments demonstrate that our framework outperforms previous image rescaling baselines in both rate-distortion performance and is much faster than prior work in HR image reconstruction speed.
****
Gated Stereo: Joint Depth Estimation From Gated and Wide-Baseline Active Stereo Cues
Stefanie Walz, Mario Bijelic, Andrea Ramazzina, Amanpreet Walia, Fahim Mannan, Felix Heide; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13252-13262
We propose Gated Stereo, a high-resolution and long-range depth estimation technique that operates on active gated stereo images. Using active and high dynamic range passive captures, Gated Stereo exploits multi-view cues alongside time-of-flight intensity cues from active gating. To this end, we propose a depth estimation method with a monocular and stereo depth prediction branch which are combined in a final fusion stage. Each block is supervised through a combination of supervised and gated self-supervision losses. To facilitate training and validation, we acquire a long-range synchronized gated stereo dataset for automotive scenarios. We find that the method achieves an improvement of more than 50 % MAE compared to the next best RGB stereo method, and 74 % MAE to existing monocular gated methods for distances up to 160 m. Our code, models and datasets are available here: https://light.princeton.edu/gatedstereo/.
****
Label Information Bottleneck for Label Enhancement
Qinghai Zheng, Jihua Zhu, Haoyu Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7497-7506
In this work, we focus on the challenging problem of Label Enhancement (LE), which aims to exactly recover label distributions from logical labels, and present a novel Label Information Bottleneck (LIB) method for LE. For the recovery process of label distributions, the label irrelevant information contained in the dataset may lead to unsatisfactory recovery performance. To address this limitation, we make efforts to excavate the essential label relevant information to improve the recovery performance. Our method formulates the LE problem as the following two joint processes: 1) learning the representation with the essential label relevant information, 2) recovering label distributions based on the learned representation. The label relevant information can be excavated based on the "bottleneck" formed by the learned representation. Significantly, both the label relevant information about the label assignments and the label relevant information about the label gaps can be explored in our method. Evaluation experiments conducted on several benchmark label distribution learning datasets verify the effectiveness and competitiveness of LIB.
****
Multi-Modal Representation Learning With Text-Driven Soft Masks
Jaeyoo Park, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2798-2807
We propose a visual-linguistic representation learning approach within a self-supervised learning framework by introducing a new operation, loss, and data augmentation strategy. First, we generate diverse features for the image-text matching (ITM) task via soft-masking the regions in an image, which are most relevant to a certain word in the corresponding caption, instead of completely removing them. Since our framework relies only on image-caption pairs with no fine-grained annotations, we identify the relevant regions to each word by computing the word-conditional visual attention using multi-modal encoder. Second, we encourage the model to focus more on hard but diverse examples by proposing a focal loss for the image-text contrastive learning (ITC) objective, which alleviates the inherent limitations of overfitting and bias issues. Last, we perform multi-modal data augmentations for self-supervised learning via mining various examples by masking texts and rendering distortions on images. We show that the combination of these three innovations is effective for learning a pretrained model, leading to outstanding performance on multiple vision-language downstream tasks.

****
Gazeformer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention
Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, Minh Hoai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1441-1450
Predicting human gaze is important in Human-Computer Interaction (HCI). However, to practically serve HCI applications, gaze prediction models must be scalable, fast, and accurate in their spatial and temporal gaze predictions. Recent scanpath prediction models focus on goal-directed attention (search). Such models are limited in their application due to a common approach relying on trained target detectors for all possible objects, and the availability of human gaze data for their training (both not scalable). In response, we pose a new task called ZeroGaze, a new variant of zero-shot learning where gaze is predicted for never-before-searched objects, and we develop a novel model, Gazeformer, to solve the ZeroGaze problem. In contrast to existing methods using object detector modules, Gazeformer encodes the target using a natural language model, thus leveraging semantic similarities in scanpath prediction. We use a transformer-based encoder-decoder architecture because transformers are particularly useful for generating contextual representations. Gazeformer surpasses other models by a large margin (19% - 70%) on the ZeroGaze setting. It also outperforms existing target-detection models on standard gaze prediction for both target-present and target-absent search tasks. In addition to its improved performance, Gazeformer is more than five times faster than the state-of-the-art target-present visual search model.
****
MammalNet: A Large-Scale Video Benchmark for Mammal Recognition and Behavior Understanding
Jun Chen, Ming Hu, Darren J. Coker, Michael L. Berumen, Blair Costelloe, Sara Beery, Anna Rohrbach, Mohamed Elhoseiny; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13052-13061
Monitoring animal behavior can facilitate conservation efforts by providing key insights into wildlife health, population status, and ecosystem function. Automatic recognition of animals and their behaviors is critical for capitalizing on the large unlabeled datasets generated by modern video devices and for accelerating monitoring efforts at scale. However, the development of automated recognition systems is currently hindered by a lack of appropriately labeled datasets. Existing video datasets 1) do not classify animals according to established biological taxonomies; 2) are too small to facilitate large-scale behavioral studies and are often limited to a single species; and 3) do not feature temporally localized annotations and therefore do not facilitate localization of targeted behaviors within longer video sequences. Thus, we propose MammalNet, a new large-scale animal behavior dataset with taxonomy-guided annotations of mammals and their common behaviors. MammalNet contains over 18K videos totaling 539 hours, which is 10 times larger than the largest existing animal behavior dataset. It covers 17 orders, 69 families, and 173 mammal categories for animal categorization and captures 12 high-level animal behaviors that received focus in previous animal behavior studies. We establish three benchmarks on MammalNet: standard animal and behavior recognition, compositional low-shot animal and behavior recognition, and behavior detection. Our dataset and code have been made available at: https://mammal-net.github.io.
****
Hand Avatar: Free-Pose Hand Animation and Rendering From Monocular Video
Xingyu Chen, Baoyuan Wang, Heung-Yeung Shum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8683-8693
We present HandAvatar, a novel representation for hand animation and rendering, which can generate smoothly compositional geometry and self-occlusion-aware texture. Specifically, we first develop a MANO-HD model as a high-resolution mesh topology to fit personalized hand shapes. Sequentially, we decompose hand geometry into per-bone rigid parts, and then re-compose paired geometry encodings to derive an across-part consistent occupancy field. As for texture modeling, we propo

se a self-occlusion-aware shading field (SelF). In SelF, drivable anchors are paved on the MANO-HD surface to record albedo information under a wide variety of hand poses. Moreover, directed soft occupancy is designed to describe the ray-to-surface relation, which is leveraged to generate an illumination field for the disentanglement of pose-independent albedo and pose-dependent illumination. Trained from monocular video data, our HandAvatar can perform free-pose hand animation and rendering while at the same time achieving superior appearance fidelity. We also demonstrate that HandAvatar provides a route for hand appearance editing.
****
Rethinking the Correlation in Few-Shot Segmentation: A Buoys View
Yuan Wang, Rui Sun, Tianzhu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7183-7192
Few-shot segmentation (FSS) aims to segment novel objects in a given query image with only a few annotated support images. However, most previous best-performing methods, whether prototypical learning methods or affinity learning methods, neglect to alleviate false matches caused by their own pixel-level correlation. In this work, we rethink how to mitigate the false matches from the perspective of representative reference features (referred to as buoys), and propose a novel adaptive buoys correlation (ABC) network to rectify direct pairwise pixel-level correlation, including a buoys mining module and an adaptive correlation module. The proposed ABC enjoys several merits. First, to learn the buoys well without any correspondence supervision, we customize the buoys mining module according to the three characteristics of representativeness, task awareness and resilience. Second, the proposed adaptive correlation module is responsible for further endowing buoy-correlation-based pixel matching with an adaptive ability. Extensive experimental results with two different backbones on two challenging benchmarks demonstrate that our ABC, as a general plugin, achieves consistent improvements over several leading methods on both 1-shot and 5-shot settings.
****
VindLU: A Recipe for Effective Video-and-Language Pretraining
Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, Gedas Bertasius; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10739-10750
The last several years have witnessed remarkable progress in video-and-language (VidL) understanding. However, most modern VidL approaches use complex and specialized model architectures and sophisticated pretraining protocols, making the reproducibility, analysis and comparisons of these frameworks difficult. Hence, instead of proposing yet another new VidL model, this paper conducts a thorough empirical study demystifying the most important factors in the VidL model design. Among the factors that we investigate are (i) the spatiotemporal architecture design, (ii) the multimodal fusion schemes, (iii) the pretraining objectives, (iv) the choice of pretraining data, (v) pretraining and finetuning protocols, and (vi) dataset and model scaling. Our empirical study reveals that the most important design factors include: temporal modeling, video-to-text multimodal fusion, masked modeling objectives, and joint training on images and videos. Using these empirical insights, we then develop a step-by-step recipe, dubbed VindLU, for effective VidL pretraining. Our final model trained using our recipe achieves comparable or better than state-of-the-art results on several VidL tasks without relying on external CLIP pretraining. In particular, on the text-to-video retrieval task, our approach obtains 61.2% on DiDeMo, and 55.0% on ActivityNet, outperforming current SOTA by 7.8% and 6.1% respectively. Furthermore, our model also obtains state-of-the-art video question-answering results on ActivityNet-QA, MSRVTT-QA, MSRVTT-MC and TVQA. Our code and pretrained models are publicly available at: https://github.com/klauscc/VindLU.
****
Scaling Language-Image Pre-Training via Masking
Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, Kaiming He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23390-23400

We present Fast Language-Image Pre-training (FLIP), a simple and more efficient method for training CLIP. Our method randomly masks out and removes a large portion of image patches during training. Masking allows us to learn from more image-text pairs given the same wall-clock time and contrast more samples per iteration with similar memory footprint. It leads to a favorable trade-off between accuracy and training time. In our experiments on 400 million image-text pairs, FLIP improves both accuracy and speed over the no-masking baseline. On a large diversity of downstream tasks, FLIP dominantly outperforms the CLIP counterparts trained on the same data. Facilitated by the speedup, we explore the scaling behavior of increasing the model size, data size, or training length, and report encouraging results and comparisons. We hope that our work will foster future research on scaling vision-language learning.
****
OmniAvatar: Geometry-Guided Controllable 3D Head Synthesis
Hongyi Xu, Guoxian Song, Zihang Jiang, Jianfeng Zhang, Yichun Shi, Jing Liu, Wanchun Ma, Jiashi Feng, Linjie Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12814-12824
We present OmniAvatar, a novel geometry-guided 3D head synthesis model trained from in-the-wild unstructured images that is capable of synthesizing diverse identity-preserved 3D heads with compelling dynamic details under full disentangled control over camera poses, facial expressions, head shapes, articulated neck and jaw poses. To achieve such high level of disentangled control, we first explicitly define a novel semantic signed distance function (SDF) around a head geometry (FLAME) conditioned on the control parameters. This semantic SDF allows us to build a differentiable volumetric correspondence map from the observation space to a disentangled canonical space from all the control parameters. We then leverage the 3D-aware GAN framework (EG3D) to synthesize detailed shape and appearance of 3D full heads in the canonical space, followed by a volume rendering step guided by the volumetric correspondence map to output into the observation space. To ensure the control accuracy on the synthesized head shapes and expressions, we introduce a geometry prior loss to conform to head SDF and a control loss to conform to the expression code. Further, we enhance the temporal realism with dynamic details conditioned upon varying expressions and joint poses. Our model can synthesize more preferable identity-preserved 3D heads with compelling dynamic details compared to the state-of-the-art methods both qualitatively and quantitatively. We also provide an ablation study to justify many of our system design choices.
****
DiffRF: Rendering-Guided 3D Radiance Field Diffusion
Norman Müller, Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, Matthias Nießner; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4328-4338
We introduce DiffRF, a novel approach for 3D radiance field synthesis based on denoising diffusion probabilistic models. While existing diffusion-based methods operate on images, latent codes, or point cloud data, we are the first to directly generate volumetric radiance fields. To this end, we propose a 3D denoising model which directly operates on an explicit voxel grid representation. However, as radiance fields generated from a set of posed images can be ambiguous and contain artifacts, obtaining ground truth radiance field samples is non-trivial. We address this challenge by pairing the denoising formulation with a rendering loss, enabling our model to learn a deviated prior that favours good image quality instead of trying to replicate fitting errors like floating artifacts. In contrast to 2D-diffusion models, our model learns multi-view consistent priors, enabling free-view synthesis and accurate shape generation. Compared to 3D GANs, our diffusion-based approach naturally enables conditional generation like masked completion or single-view 3D synthesis at inference time.
****
DNF: Decouple and Feedback Network for Seeing in the Dark
Xin Jin, Ling-Hao Han, Zhen Li, Chun-Le Guo, Zhi Chai, Chongyi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202

3, pp. 18135-18144
The exclusive properties of RAW data have shown great potential for low-light im age enhancement. Nevertheless, the performance is bottlenecked by the inherent l imitations of existing architectures in both single-stage and multi-stage method s. Mixed mapping across two different domains, noise-to-clean and RAW-to-sRGB, m isleads the single-stage methods due to the domain ambiguity. The multi-stage me thods propagate the information merely through the resulting image of each stage , neglecting the abundant features in the lossy image-level dataflow. In this pa per, we probe a generalized solution to these bottlenecks and propose a Decouple aNd Feedback framework, abbreviated as DNF. To mitigate the domain ambiguity, d omainspecific subtasks are decoupled, along with fully utilizing the unique prop erties in RAW and sRGB domains. The feature propagation across stages with a fee dback mechanism avoids the information loss caused by image-level dataflow. The two key insights of our method resolve the inherent limitations of RAW data-base d low-light image enhancement satisfactorily, empowering our method to outperfor m the previous state-of-the-art method by a large margin with only 19% parameter s, achieving 0.97dB and 1.30dB PSNR improvements on the Sony and Fuji subsets of SID.
****
SUDS: Scalable Urban Dynamic Scenes
Haithem Turki, Jason Y. Zhang, Francesco Ferroni, Deva Ramanan; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12375-12385
We extend neural radiance fields (NeRFs) to dynamic large-scale urban scenes. Pr ior work tends to reconstruct single video clips of short durations (up to 10 se conds). Two reasons are that such methods (a) tend to scale linearly with the nu mber of moving objects and input videos because a separate model is built for ea ch and (b) tend to require supervision via 3D bounding boxes and panoptic labels , obtained manually or via category-specific models. As a step towards truly ope n-world reconstructions of dynamic cities, we introduce two key innovations: (a) we factorize the scene into three separate hash table data structures to effici ently encode static, dynamic, and far-field radiance fields, and (b) we make use of unlabeled target signals consisting of RGB images, sparse LiDAR, off-the-she lf self-supervised 2D descriptors, and most importantly, 2D optical flow. Operat ionalizing such inputs via photometric, geometric, and feature-metric reconstruc tion losses enables SUDS to decompose dynamic scenes into the static background, individual objects, and their motions. When combined with our multi-branch tabl e representation, such reconstructions can be scaled to tens of thousands of obj ects across 1.2 million frames from 1700 videos spanning geospatial footprints o f hundreds of kilometers, (to our knowledge) the largest dynamic NeRF built to d ate. We present qualitative initial results on a variety of tasks enabled by our representations, including novel-view synthesis of dynamic urban scenes, unsupe rvised 3D instance segmentation, and unsupervised 3D cuboid detection. To compar e to prior work, we also evaluate on KITTI and Virtual KITTI 2, surpassing state -of-the-art methods that rely on ground truth 3D bounding box annotations while being 10x quicker to train.
****
Deformable Mesh Transformer for 3D Human Mesh Recovery
Yusuke Yoshiyasu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17006-17015
We present Deformable mesh transFormer (DeFormer), a novel vertex-based approach to monocular 3D human mesh recovery. DeFormer iteratively fits a body mesh mode l to an input image via a mesh alignment feedback loop formed within a transform er decoder that is equipped with efficient body mesh driven attention modules: 1) body sparse self-attention and 2) deformable mesh cross attention. As a result , DeFormer can effectively exploit high-resolution image feature maps and a dens e mesh model which were computationally expensive to deal with in previous appro aches using the standard transformer attention. Experimental results show that D eFormer achieves state-of-the-art performances on the Human3.6M and 3DPW benchma rks. Ablation study is also conducted to show the effectiveness of the DeFormer

model designs for leveraging multi-scale feature maps. Code is available at http
s://github.com/yusukey03012/DeFormer.
****
Vita-CLIP: Video and Text Adaptive CLIP via Multimodal Prompting
Syed Talal Wasim, Muzammal Naseer, Salman Khan, Fahad Shahbaz Khan, Mubarak Shah
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2023, pp. 23034-23044
Adopting contrastive image-text pretrained models like CLIP towards video classi
fication has gained attention due to its cost-effectiveness and competitive perf
ormance. However, recent works in this area face a trade-off. Finetuning the pre
trained model to achieve strong supervised performance results in low zero-shot
generalization. Similarly, freezing the backbone to retain zero-shot capability
causes significant drop in supervised accuracy. Because of this, recent works in
 literature typically train separate models for supervised and zero-shot action
recognition. In this work, we propose a multimodal prompt learning scheme that w
orks to balance the supervised and zero-shot performance under a single unified
training. Our prompting approach on the vision side caters for three aspects: 1)
 Global video-level prompts to model the data distribution; 2) Local frame-level
 prompts to provide per-frame discriminative conditioning; and 3) a summary prom
pt to extract a condensed video representation. Additionally, we define a prompt
ing scheme on the text side to augment the textual context. Through this prompti
ng scheme, we can achieve state-of-the-art zero-shot performance on Kinetics-600
, HMDB51 and UCF101 while remaining competitive in the supervised setting. By ke
eping the pretrained backbone frozen, we optimize a much lower number of paramet
ers and retain the existing general representation which helps achieve the stron
g zero-shot performance. Our codes and models will be publicly released.
****
HS-Pose: Hybrid Scope Feature Extraction for Category-Level Object Pose Estimati
on
Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis,
Wei Zhang, Hyung Jin Chang; Proceedings of the IEEE/CVF Conference on Computer V
ision and Pattern Recognition (CVPR), 2023, pp. 17163-17173
In this paper, we focus on the problem of category-level object pose estimation,
 which is challenging due to the large intra-category shape variation. 3D graph
convolution (3D-GC) based methods have been widely used to extract local geometr
ic features, but they have limitations for complex shaped objects and are sensit
ive to noise. Moreover, the scale and translation invariant properties of 3D-GC
restrict the perception of an object's size and translation information. In this
 paper, we propose a simple network structure, the HS-layer, which extends 3D-GC
 to extract hybrid scope latent features from point cloud data for category-leve
l object pose estimation tasks. The proposed HS-layer: 1) is able to perceive lo
cal-global geometric structure and global information, 2) is robust to noise, an
d 3) can encode size and translation information. Our experiments show that the
simple replacement of the 3D-GC layer with the proposed HS-layer on the baseline
 method (GPV-Pose) achieves a significant improvement, with the performance incr
eased by 14.5% on 5d2cm metric and 10.3% on IoU75. Our method outperforms the st
ate-of-the-art methods by a large margin (8.3% on 5d2cm, 6.9% on IoU75) on REAL2
75 dataset and runs in real-time (50 FPS).
****
Cloud-Device Collaborative Adaptation to Continual Changing Environments in the
Real-World
Yulu Gan, Mingjie Pan, Rongyu Zhang, Zijian Ling, Lingran Zhao, Jiaming Liu, Sha
nghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Patt
ern Recognition (CVPR), 2023, pp. 12157-12166
When facing changing environments in the real world, the lightweight model on cl
ient devices suffer from severe performance drop under distribution shifts. The
main limitations of existing device model lie in: (1) unable to update due to th
e computation limit of the device, (2) limited generalization ability of the lig
htweight model. Meanwhile, recent large models have shown strong generalization
capability on cloud while they can not be deployed on client devices due to the

poor computation constraint. To enable the device model to deal with changing en vironments, we propose a new learning paradigm of Cloud-Device Collaborative Con tinual Adaptation. To encourage collaboration between cloud and device and impro ve the generalization of device model, we propose an Uncertainty-based Visual Pr ompt Adapted (U-VPA) teacher-student model in such paradigm. Specifically, we fi rst design the Uncertainty Guided Sampling (UGS) to screen out challenging data continuously and transmit the most out-of-distribution samples from the device t o the cloud. To further transfer the generalization capability of the large mode l on the cloud to the device model, we propose a Visual Prompt Learning Strategy with Uncertainty guided updating (VPLU) to specifically deal with the selected samples with more distribution shifts. Then, we transmit the visual prompts to t he device and concatenate them with the incoming data to pull the device testing distribution closer to the cloud training distribution. We conduct extensive ex periments on two object detection datasets with continually changing environment s. Our proposed U-VPA teacher-student framework outperforms previous state-of-th e-art test time adaptation and device-cloud collaboration methods. The code and datasets will be released.
****
Parts2Words: Learning Joint Embedding of Point Clouds and Texts by Bidirectional Matching Between Parts and Words
Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, Yi Chang; Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6884 -6893
Shape-Text matching is an important task of high-level shape understanding. Curr ent methods mainly represent a 3D shape as multiple 2D rendered views, which obv iously can not be understood well due to the structural ambiguity caused by self -occlusion in the limited number of views. To resolve this issue, we directly re present 3D shapes as point clouds, and propose to learn joint embedding of point clouds and texts by bidirectional matching between parts from shapes and words from texts. Specifically, we first segment the point clouds into parts, and then leverage optimal transport method to match parts and words in an optimized feat ure space, where each part is represented by aggregating features of all points within it and each word is abstracted by its contextual information. We optimize the feature space in order to enlarge the similarities between the paired train ing samples, while simultaneously maximizing the margin between the unpaired one s. Experiments demonstrate that our method achieves a significant improvement in accuracy over the SOTAs on multi-modal retrieval tasks under the Text2Shape dat aset. Codes are available at https://github.com/JLUtangchuan/Parts2Words.
****
Proposal-Based Multiple Instance Learning for Weakly-Supervised Temporal Action Localization
Huan Ren, Wenfei Yang, Tianzhu Zhang, Yongdong Zhang; Proceedings of the IEEE/CV F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2394-2 404
Weakly-supervised temporal action localization aims to localize and recognize ac tions in untrimmed videos with only video-level category labels during training. Without instance-level annotations, most existing methods follow the Segment-ba sed Multiple Instance Learning (S-MIL) framework, where the predictions of segme nts are supervised by the labels of videos. However, the objective for acquiring segment-level scores during training is not consistent with the target for acqu iring proposal-level scores during testing, leading to suboptimal results. To de al with this problem, we propose a novel Proposal-based Multiple Instance Learni ng (P-MIL) framework that directly classifies the candidate proposals in both th e training and testing stages, which includes three key designs: 1) a surroundin g contrastive feature extraction module to suppress the discriminative short pro posals by considering the surrounding contrastive information, 2) a proposal com pleteness evaluation module to inhibit the low-quality proposals with the guidan ce of the completeness pseudo labels, and 3) an instance-level rank consistency loss to achieve robust detection by leveraging the complementarity of RGB and FL OW modalities. Extensive experimental results on two challenging benchmarks incl

uding THUMOS14 and ActivityNet demonstrate the superior performance of our metho
d. Our code is available at github.com/OpenSpaceAI/CVPR2023_P-MIL.
****
LayoutDM: Transformer-Based Diffusion Model for Layout Generation
Shang Chai, Liansheng Zhuang, Fengying Yan; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18349-18358
Automatic layout generation that can synthesize high-quality layouts is an impor
tant tool for graphic design in many applications. Though existing methods based
 on generative models such as Generative Adversarial Networks (GANs) and Variati
onal Auto-Encoders (VAEs) have progressed, they still leave much room for improv
ing the quality and diversity of the results. Inspired by the recent success of
diffusion models in generating high-quality images, this paper explores their po
tential for conditional layout generation and proposes Transformer-based Layout
Diffusion Model (LayoutDM) by instantiating the conditional denoising diffusion
probabilistic model (DDPM) with a purely transformer-based architecture. Instead
 of using convolutional neural networks, a transformer-based conditional Layout
Denoiser is proposed to learn the reverse diffusion process to generate samples
from noised layout data. Benefitting from both transformer and DDPM, our LayoutD
M is of desired properties such as high-quality generation, strong sample divers
ity, faithful distribution coverage, and stationary training in comparison to GA
Ns and VAEs. Quantitative and qualitative experimental results show that our met
hod outperforms state-of-the-art generative models in terms of quality and diver
sity.
****
HandNeRF: Neural Radiance Fields for Animatable Interacting Hands
Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, Houqiang Li; Proceedings of the IEEE
/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 210
78-21087
We propose a novel framework to reconstruct accurate appearance and geometry wit
h neural radiance fields (NeRF) for interacting hands, enabling the rendering of
 photo-realistic images and videos for gesture animation from arbitrary views. G
iven multi-view images of a single hand or interacting hands, an off-the-shelf s
keleton estimator is first employed to parameterize the hand poses. Then we desi
gn a pose-driven deformation field to establish correspondence from those differ
ent poses to a shared canonical space, where a pose-disentangled NeRF for one ha
nd is optimized. Such unified modeling efficiently complements the geometry and
texture cues in rarely-observed areas for both hands. Meanwhile, we further leve
rage the pose priors to generate pseudo depth maps as guidance for occlusion-awa
re density learning. Moreover, a neural feature distillation method is proposed
to achieve cross-domain alignment for color optimization. We conduct extensive e
xperiments to verify the merits of our proposed HandNeRF and report a series of
state-of-the-art results both qualitatively and quantitatively on the large-scal
e InterHand2.6M dataset.
****
ASPnet: Action Segmentation With Shared-Private Representation of Multiple Data
Sources
Beatrice van Amsterdam, Abdolrahim Kadkhodamohammadi, Imanol Luengo, Danail Stoy
anov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2023, pp. 2384-2393
Most state-of-the-art methods for action segmentation are based on single input
modalities or naive fusion of multiple data sources. However, effective fusion o
f complementary information can potentially strengthen segmentation models and m
ake them more robust to sensor noise and more accurate with smaller training dat
asets. In order to improve multimodal representation learning for action segment
ation, we propose to disentangle hidden features of a multi-stream segmentation
model into modality-shared components, containing common information across data
 sources, and private components; we then use an attention bottleneck to capture
 long-range temporal dependencies in the data while preserving disentanglement i
n consecutive processing layers. Evaluation on 50salads, Breakfast and RARP45 da
tasets shows that our multimodal approach outperforms different data fusion base

lines on both multiview and multimodal data sources, obtaining competitive or be
tter results compared with the state-of-the-art. Our model is also more robust t
o additive sensor noise and can achieve performance on par with strong video bas
elines even with less training data.
****
Seasoning Model Soups for Robustness to Adversarial and Natural Distribution Shi
fts
Francesco Croce, Sylvestre-Alvise Rebuffi, Evan Shelhamer, Sven Gowal; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2023, pp. 12313-12323
Adversarial training is widely used to make classifiers robust to a specific thr
eat or adversary, such as l_p-norm bounded perturbations of a given p-norm. Howe
ver, existing methods for training classifiers robust to multiple threats requir
e knowledge of all attacks during training and remain vulnerable to unseen distr
ibution shifts. In this work, we describe how to obtain adversarially-robust mod
el soups (i.e., linear combinations of parameters) that smoothly trade-off robus
tness to different l_p-norm bounded adversaries. We demonstrate that such soups
allow us to control the type and level of robustness, and can achieve robustness
 to all threats without jointly training on all of them. In some cases, the resu
lting model soups are more robust to a given l_p-norm adversary than the constit
uent model specialized against that same adversary. Finally, we show that advers
arially-robust model soups can be a viable tool to adapt to distribution shifts
from a few examples.
****
Introducing Competition To Boost the Transferability of Targeted Adversarial Exa
mples Through Clean Feature Mixup
Junyoung Byun, Myung-Joon Kwon, Seungju Cho, Yoonji Kim, Changick Kim; Proceedin
gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),
 2023, pp. 24648-24657
Deep neural networks are widely known to be susceptible to adversarial examples,
 which can cause incorrect predictions through subtle input modifications. These
 adversarial examples tend to be transferable between models, but targeted attac
ks still have lower attack success rates due to significant variations in decisi
on boundaries. To enhance the transferability of targeted adversarial examples,
we propose introducing competition into the optimization process. Our idea is to
 craft adversarial perturbations in the presence of two new types of competitor
noises: adversarial perturbations towards different target classes and friendly
perturbations towards the correct class. With these competitors, even if an adve
rsarial example deceives a network to extract specific features leading to the t
arget class, this disturbance can be suppressed by other competitors. Therefore,
 within this competition, adversarial examples should take different attack stra
tegies by leveraging more diverse features to overwhelm their interference, lead
ing to improving their transferability to different models. Considering the comp
utational complexity, we efficiently simulate various interference from these tw
o types of competitors in feature space by randomly mixing up stored clean featu
res in the model inference and named this method Clean Feature Mixup (CFM). Our
extensive experimental results on the ImageNet-Compatible and CIFAR-10 datasets
show that the proposed method outperforms the existing baselines with a clear ma
rgin. Our code is available at https://github.com/dreamflake/CFM.
****
Ingredient-Oriented Multi-Degradation Learning for Image Restoration
Jinghao Zhang, Jie Huang, Mingde Yao, Zizheng Yang, Hu Yu, Man Zhou, Feng Zhao;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2023, pp. 5825-5835
Learning to leverage the relationship among diverse image restoration tasks is q
uite beneficial for unraveling the intrinsic ingredients behind the degradation.
 Recent years have witnessed the flourish of various All-in-one methods, which h
andle multiple image degradations within a single model. In practice, however, f
ew attempts have been made to excavate task correlations in that exploring the u
nderlying fundamental ingredients of various image degradations, resulting in po

or scalability as more tasks are involved. In this paper, we propose a novel per
spective to delve into the degradation via an ingredients-oriented rather than p
revious task-oriented manner for scalable learning. Specifically, our method, na
med Ingredients-oriented Degradation Reformulation framework (IDR), consists of
two stages, namely task-oriented knowledge collection and ingredients-oriented k
nowledge integration. In the first stage, we conduct ad hoc operations on differ
ent degradations according to the underlying physics principles, and establish t
he corresponding prior hubs for each type of degradation. While the second stage
 progressively reformulates the preceding task-oriented hubs into single ingredi
ents-oriented hub via learnable Principal Component Analysis (PCA), and employs
a dynamic routing mechanism for probabilistic unknown degradation removal. Exten
sive experiments on various image restoration tasks demonstrate the effectivenes
s and scalability of our method. More importantly, our IDR exhibits the favorabl
e generalization ability to unknown downstream tasks.
****
How To Prevent the Continuous Damage of Noises To Model Training?
Xiaotian Yu, Yang Jiang, Tianqi Shi, Zunlei Feng, Yuexuan Wang, Mingli Song, Li
Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog
nition (CVPR), 2023, pp. 12054-12063
Deep learning with noisy labels is challenging and inevitable in many circumstan
ces. Existing methods reduce the impact of noise samples by reducing loss weight
s of uncertain samples or by filtering out potential noise samples, which highly
 rely on the model's superior discriminative power for identifying noise samples
. However, in the training stage, the trainee model is imperfect will miss many
noise samples, which cause continuous damage to the model training. Consequently
, there is a large performance gap between existing anti-noise models trained wi
th noisy samples and models trained with clean samples. In this paper, we put fo
rward a Gradient Switching Strategy (GSS) to prevent the continuous damage of no
ise samples to the classifier. Theoretical analysis shows that the damage comes
from the misleading gradient direction computed from the noise samples. The trai
nee model will deviate from the correct optimization direction under the influen
ce of the accumulated misleading gradient of noise samples. To address this prob
lem, the proposed GSS alleviates the damage by switching the current gradient di
rection of each sample to a new direction selected from a gradient direction poo
l, which contains all-class gradient directions with different probabilities. Du
ring training, the trainee model is optimized along switched gradient directions
 generated by GSS, which assigns higher probabilities to potential principal dir
ections for high-confidence samples. Conversely, uncertain samples have a relati
vely uniform probability distribution for all gradient directions, which can can
cel out the misleading gradient directions. Extensive experiments show that a mo
del trained with GSS can achieve comparable performance with a model trained wit
h clean data. Moreover, the proposed GSS is pluggable for existing frameworks fo
r noisy-label learning. This work can provide a new perspective for future noisy
-label learning.
****
A Whac-a-Mole Dilemma: Shortcuts Come in Multiples Where Mitigating One Amplifie
s Others
Zhiheng Li, Ivan Evtimov, Albert Gordo, Caner Hazirbas, Tal Hassner, Cristian Ca
nton Ferrer, Chenliang Xu, Mark Ibrahim; Proceedings of the IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20071-20082
Machine learning models have been found to learn shortcuts---unintended decision
 rules that are unable to generalize---undermining models' reliability. Previous
 works address this problem under the tenuous assumption that only a single shor
tcut exists in the training data. Real-world images are rife with multiple visua
l cues from background to texture. Key to advancing the reliability of vision sy
stems is understanding whether existing methods can overcome multiple shortcuts
or struggle in a Whac-A-Mole game, i.e., where mitigating one shortcut amplifies
 reliance on others. To address this shortcoming, we propose two benchmarks: 1)
UrbanCars, a dataset with precisely controlled spurious cues, and 2) ImageNet-W,
 an evaluation set based on ImageNet for watermark, a shortcut we discovered aff

ects nearly every modern vision model. Along with texture and background, ImageNet-W allows us to study multiple shortcuts emerging from training on natural images. We find computer vision models, including large foundation models---regardless of training set, architecture, and supervision---struggle when multiple shortcuts are present. Even methods explicitly designed to combat shortcuts struggle in a Whac-A-Mole dilemma. To tackle this challenge, we propose Last Layer Ensemble, a simple-yet-effective method to mitigate multiple shortcuts without Whac-A-Mole behavior. Our results surface multi-shortcut mitigation as an overlooked challenge critical to advancing the reliability of vision systems. The datasets and code are released: https://github.com/facebookresearch/Whac-A-Mole.
****
Skinned Motion Retargeting With Residual Perception of Motion Semantics & Geometry
Jiaxu Zhang, Junwu Weng, Di Kang, Fang Zhao, Shaoli Huang, Xuefei Zhe, Linchao Bao, Ying Shan, Jue Wang, Zhigang Tu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13864-13872
A good motion retargeting cannot be reached without reasonable consideration of source-target differences on both the skeleton and shape geometry levels. In this work, we propose a novel Residual RETargeting network (R2ET) structure, which relies on two neural modification modules, to adjust the source motions to fit the target skeletons and shapes progressively. In particular, a skeleton-aware module is introduced to preserve the source motion semantics. A shape-aware module is designed to perceive the geometries of target characters to reduce interpenetration and contact-missing. Driven by our explored distance-based losses that explicitly model the motion semantics and geometry, these two modules can learn residual motion modifications on the source motion to generate plausible retargeted motion in a single inference without post-processing. To balance these two modifications, we further present a balancing gate to conduct linear interpolation between them. Extensive experiments on the public dataset Mixamo demonstrate that our R2ET achieves the state-of-the-art performance, and provides a good balance between the preservation of motion semantics as well as the attenuation of interpenetration and contact-missing. Code is available at https://github.com/Kebii/R2ET.
****
Weakly-Supervised Single-View Image Relighting
Renjiao Yi, Chenyang Zhu, Kai Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8402-8411
We present a learning-based approach to relight a single image of Lambertian and low-frequency specular objects. Our method enables inserting objects from photographs into new scenes and relighting them under the new environment lighting, which is essential for AR applications. To relight the object, we solve both inverse rendering and re-rendering. To resolve the ill-posed inverse rendering, we propose a weakly-supervised method by a low-rank constraint. To facilitate the weakly-supervised training, we contribute Relit, a large-scale (750K images) dataset of videos with aligned objects under changing illuminations. For re-rendering, we propose a differentiable specular rendering layer to render low-frequency non-Lambertian materials under various illuminations of spherical harmonics. The whole pipeline is end-to-end and efficient, allowing for a mobile app implementation of AR object insertion. Extensive evaluations demonstrate that our method achieves state-of-the-art performance. Project page: https://renjiaoyi.github.io/relighting/.
****
DualVector: Unsupervised Vector Font Synthesis With Dual-Part Representation
Ying-Tian Liu, Zhifei Zhang, Yuan-Chen Guo, Matthew Fisher, Zhaowen Wang, Song-Hai Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14193-14202
Automatic generation of fonts can be an important aid to typeface design. Many current approaches regard glyphs as pixelated images, which present artifacts when scaling and inevitable quality losses after vectorization. On the other hand, existing vector font synthesis methods either fail to represent the shape concis

ely or require vector supervision during training. To push the quality of vector font synthesis to the next level, we propose a novel dual-part representation for vector glyphs, where each glyph is modeled as a collection of closed "positive" and "negative" path pairs. The glyph contour is then obtained by boolean operations on these paths. We first learn such a representation only from glyph images and devise a subsequent contour refinement step to align the contour with an image representation to further enhance details. Our method, named DualVector, outperforms state-of-the-art methods in vector font synthesis both quantitatively and qualitatively. Our synthesized vector fonts can be easily converted to common digital font formats like TrueType Font for practical use. The code is released at https://github.com/thuliu-yt16/dualvector.
****
Efficient Scale-Invariant Generator With Column-Row Entangled Pixel Synthesis
Thuan Hoang Nguyen, Thanh Van Le, Anh Tran; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22408-22417
Any-scale image synthesis offers an efficient and scalable solution to synthesize photo-realistic images at any scale, even going beyond 2K resolution. However, existing GAN-based solutions depend excessively on convolutions and a hierarchical architecture, which introduce inconsistency and the "texture sticking" issue when scaling the output resolution. From another perspective, INR-based generators are scale-equivariant by design, but their huge memory footprint and slow inference hinder these networks from being adopted in large-scale or real-time systems. In this work, we propose Column-Row Entangled Pixel Synthesisthes (CREPS), a new generative model that is both efficient and scale-equivariant without using any spatial convolutions or coarse-to-fine design. To save memory footprint and make the system scalable, we employ a novel bi-line representation that decomposes layer-wise feature maps into separate "thick" column and row encodings. Experiments on standard datasets, including FFHQ, LSUN-Church, and MetFaces, confirm CREPS' ability to synthesize scale-consistent and alias-free images up to 4K resolution with proper training and inference speed.
****
ReasonNet: End-to-End Driving With Temporal and Global Reasoning
Hao Shao, Letian Wang, Ruobing Chen, Steven L. Waslander, Hongsheng Li, Yu Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13723-13733
The large-scale deployment of autonomous vehicles is yet to come, and one of the major remaining challenges lies in urban dense traffic scenarios. In such cases, it remains challenging to predict the future evolution of the scene and future behaviors of objects, and to deal with rare adverse events such as the sudden appearance of occluded objects. In this paper, we present ReasonNet, a novel end-to-end driving framework that extensively exploits both temporal and global information of the driving scene. By reasoning on the temporal behavior of objects, our method can effectively process the interactions and relationships among features in different frames. Reasoning about the global information of the scene can also improve overall perception performance and benefit the detection of adverse events, especially the anticipation of potential danger from occluded objects. For comprehensive evaluation on occlusion events, we also release publicly a driving simulation benchmark DriveOcclusionSim consisting of diverse occlusion events. We conduct extensive experiments on multiple CARLA benchmarks, where our model outperforms all prior methods, ranking first on the sensor track of the public CARLA Leaderboard.
****
Learning Situation Hyper-Graphs for Video Question Answering
Aisha Urooj, Hilde Kuehne, Bo Wu, Kim Chheu, Walid Bousselham, Chuang Gan, Niels Lobo, Mubarak Shah; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14879-14889
Answering questions about complex situations in videos requires not only capturing of the presence of actors, objects, and their relations, but also the evolution of these relationships over time. A situation hyper-graph is a representation that describes situations as scene sub-graphs for video frames and hyper-edges

for connected sub-graphs, and has been proposed to capture all such information in a compact structured form. In this work, we propose an architecture for Video Question Answering (VQA) that enables answering questions related to video content by predicting situation hyper-graphs, coined Situation Hyper-Graph based Video Question Answering (SHG-VQA). To this end, we train a situation hyper-graph decoder to implicitly identify graph representations with actions and object/human-object relationships from the input video clip and to use cross-attention between the predicted situation hyper-graphs and the question embedding to predict the correct answer. The proposed method is trained in an end-to-end manner and optimized by a cross-entropy based VQA loss function and a Hungarian matching loss for the situation graph prediction. The effectiveness of the proposed architecture is extensively evaluated on two challenging benchmarks: AGQA and STAR. Our results show that learning the underlying situation hyper-graphs helps the system to significantly improve its performance for novel challenges of video question answering task.
****
H2ONet: Hand-Occlusion-and-Orientation-Aware Network for Real-Time 3D Hand Mesh Reconstruction
Hao Xu, Tianyu Wang, Xiao Tang, Chi-Wing Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17048-17058
Real-time 3D hand mesh reconstruction is challenging, especially when the hand is holding some object. Beyond the previous methods, we design H2ONet to fully exploit non-occluded information from multiple frames to boost the reconstruction quality. First, we decouple hand mesh reconstruction into two branches, one to exploit finger-level non-occluded information and the other to exploit global hand orientation, with lightweight structures to promote real-time inference. Second, we propose finger-level occlusion-aware feature fusion, leveraging predicted finger-level occlusion information as guidance to fuse finger-level information across time frames. Further, we design hand-level occlusion-aware feature fusion to fetch non-occluded information from nearby time frames. We conduct experiments on the Dex-YCB and HO3D-v2 datasets with challenging hand-object occlusion cases, manifesting that H2ONet is able to run in real-time and achieves state-of-the-art performance on both the hand mesh and pose precision. The code will be released on GitHub.
****
Interventional Bag Multi-Instance Learning on Whole-Slide Pathological Images
Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, Chang-Wen Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19830-19839
Multi-instance learning (MIL) is an effective paradigm for whole-slide pathological images (WSIs) classification to handle the gigapixel resolution and slide-level label. Prevailing MIL methods primarily focus on improving the feature extractor and aggregator. However, one deficiency of these methods is that the bag contextual prior may trick the model into capturing spurious correlations between bags and labels. This deficiency is a confounder that limits the performance of existing MIL methods. In this paper, we propose a novel scheme, Interventional Bag Multi-Instance Learning (IBMIL), to achieve deconfounded bag-level prediction. Unlike traditional likelihood-based strategies, the proposed scheme is based on the backdoor adjustment to achieve the interventional training, thus is capable of suppressing the bias caused by the bag contextual prior. Note that the principle of IBMIL is orthogonal to existing bag MIL methods. Therefore, IBMIL is able to bring consistent performance boosting to existing schemes, achieving new state-of-the-art performance. Code is available at https://github.com/HHHedo/IBMIL.
****
GazeNeRF: 3D-Aware Gaze Redirection With Neural Radiance Fields
Alessandro Ruzzi, Xiangwei Shi, Xi Wang, Gengyan Li, Shalini De Mello, Hyung Jin Chang, Xucong Zhang, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9676-9685
We propose GazeNeRF, a 3D-aware method for the task of gaze redirection. Existin

g gaze redirection methods operate on 2D images and struggle to generate 3D cons
istent results. Instead, we build on the intuition that the face region and eye
balls are separate 3D structures that move in a coordinated yet independent fash
ion. Our method leverages recent advancements in conditional image-based neural
radiance fields and proposes a two-branch architecture that predicts volumetric
features for the face and eye regions separately. Rigidly transforming the eye f
eatures via a 3D rotation matrix provides fine-grained control over the desired
gaze angle. The final, redirected image is then attained via differentiable volu
me compositing. Our experiments show that this architecture outperforms naively
conditioned NeRF baselines as well as previous state-of-the-art 2D gaze redirect
ion methods in terms of redirection accuracy and identity preservation. Code and
models will be released for research purposes.
****
How Can Objects Help Action Recognition?
Xingyi Zhou, Anurag Arnab, Chen Sun, Cordelia Schmid; Proceedings of the IEEE/CV
F Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2353-2
362
Current state-of-the-art video models process a video clip as a long sequence of
spatio-temporal tokens. However, they do not explicitly model objects, their in
teractions across the video, and instead process all the tokens in the video. In
this paper, we investigate how we can use knowledge of objects to design better
video models, namely to process fewer tokens and to improve recognition accurac
y. This is in contrast to prior works which either drop tokens at the cost of ac
curacy, or increase accuracy whilst also increasing the computation required. Fi
rst, we propose an object-guided token sampling strategy that enables us to reta
in a small fraction of the input tokens with minimal impact on accuracy. And sec
ond, we propose an object-aware attention module that enriches our feature repre
sentation with object information and improves overall accuracy. Our resulting f
ramework achieves better performance when using fewer tokens than strong baselin
es. In particular, we match our baseline with 30%, 40%, and 60% of the input tok
ens on SomethingElse, Something-something v2, and Epic-Kitchens, respectively. W
hen we use our model to process the same number of tokens as our baseline, we im
prove by 0.6 to 4.2 points on these datasets.
****
Realistic Saliency Guided Image Enhancement
S. Mahdi H. Miangoleh, Zoya Bylinskii, Eric Kee, Eli Shechtman, Ya■iz Aksoy; Pro
ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (
CVPR), 2023, pp. 186-194
Common editing operations performed by professional photographers include the cl
eanup operations: de-emphasizing distracting elements and enhancing subjects. Th
ese edits are challenging, requiring a delicate balance between manipulating the
viewer's attention while maintaining photo realism. While recent approaches can
boast successful examples of attention attenuation or amplification, most of th
em also suffer from frequent unrealistic edits. We propose a realism loss for sa
liency-guided image enhancement to maintain high realism across varying image ty
pes, while attenuating distractors and amplifying objects of interest. Evaluatio
ns with professional photographers confirm that we achieve the dual objective of
realism and effectiveness, and outperform the recent approaches on their own da
tasets, while requiring a smaller memory footprint and runtime. We thus offer a
viable solution for automating image enhancement and photo cleanup operations.
****
SLOPER4D: A Scene-Aware Dataset for Global 4D Human Pose Estimation in Urban Env
ironments
Yudi Dai, Yitai Lin, Xiping Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yue
xin Ma, Cheng Wang; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2023, pp. 682-692
We present SLOPER4D, a novel scene-aware dataset collected in large urban enviro
nments to facilitate the research of global human pose estimation (GHPE) with hu
man-scene interaction in the wild. Employing a head-mounted device integrated wi
th a LiDAR and camera, we record 12 human subjects' activities over 10 diverse u

rban scenes from an egocentric view. Frame-wise annotations for 2D key points, 3D pose parameters, and global translations are provided, together with reconstructed scene point clouds. To obtain accurate 3D ground truth in such large dynamic scenes, we propose a joint optimization method to fit local SMPL meshes to the scene and fine-tune the camera calibration during dynamic motions frame by frame, resulting in plausible and scene-natural 3D human poses. Eventually, SLOPER4D consists of 15 sequences of human motions, each of which has a trajectory length of more than 200 meters (up to 1,300 meters) and covers an area of more than 200 square meters (up to 30,000 square meters), including more than 100k LiDAR frames, 300k video frames, and 500K IMU-based motion frames. With SLOPER4D, we provide a detailed and thorough analysis of two critical tasks, including camera-based 3D HPE and LiDAR-based 3D HPE in urban environments, and benchmark a new task, GHPE. The in-depth analysis demonstrates SLOPER4D poses significant challenges to existing methods and produces great research opportunities. The dataset and code are released at http://www.lidarhumanmotion.net/sloper4d/.
****
SegLoc: Learning Segmentation-Based Representations for Privacy-Preserving Visual Localization
Maxime Pietrantoni, Martin Humenberger, Torsten Sattler, Gabriela Csurka; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15380-15391
Inspired by properties of semantic segmentation, in this paper we investigate how to leverage robust image segmentation in the context of privacy-preserving visual localization. We propose a new localization framework, SegLoc, that leverages image segmentation to create robust, compact, and privacy-preserving scene representations, i.e., 3D maps. We build upon the correspondence-supervised, fine-grained segmentation approach from Larsson et al (ICCV'19), making it more robust by learning a set of cluster labels with discriminative clustering, additional consistency regularization terms and we jointly learn a global image representation along with a dense local representation. In our localization pipeline, the former will be used for retrieving the most similar images, the latter to refine the retrieved poses by minimizing the label inconsistency between the 3D points of the map and their projection onto the query image. In various experiments, we show that our proposed representation allows to achieve (close-to) state-of-the-art pose estimation results while only using a compact 3D map that does not contain enough information about the original images for an attacker to reconstruct personal information.
****
Efficient Hierarchical Entropy Model for Learned Point Cloud Compression
Rui Song, Chunyang Fu, Shan Liu, Ge Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14368-14377
Learning an accurate entropy model is a fundamental way to remove the redundancy in point cloud compression. Recently, the octree-based auto-regressive entropy model which adopts the self-attention mechanism to explore dependencies in a large-scale context is proved to be promising. However, heavy global attention computations and auto-regressive contexts are inefficient for practical applications. To improve the efficiency of the attention model, we propose a hierarchical attention structure that has a linear complexity to the context scale and maintains the global receptive field. Furthermore, we present a grouped context structure to address the serial decoding issue caused by the auto-regression while preserving the compression performance. Experiments demonstrate that the proposed entropy model achieves superior rate-distortion performance and significant decoding latency reduction compared with the state-of-the-art large-scale auto-regressive entropy model.
****
RankMix: Data Augmentation for Weakly Supervised Learning of Classifying Whole Slide Images With Diverse Sizes and Imbalanced Categories
Yuan-Chih Chen, Chun-Shien Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23936-23945
Whole Slide Images (WSIs) are usually gigapixel in size and lack pixel-level ann

otations. The WSI datasets are also imbalanced in categories. These unique chara cteristics, significantly different from the ones in natural images, pose the ch allenge of classifying WSI images as a kind of weakly supervise learning problem s. In this study, we propose, RankMix, a data augmentation method of mixing rank ed features in a pair of WSIs. RankMix introduces the concepts of pseudo labelin g and ranking in order to extract key WSI regions in contributing to the WSI cla ssification task. A two-stage training is further proposed to boost stable train ing and model performance. To our knowledge, the study of weakly supervised lear ning from the perspective of data augmentation to deal with the WSI classificati on problem that suffers from lack of training data and imbalance of categories i s relatively unexplored.
****
ActMAD: Activation Matching To Align Distributions for Test-Time-Training
Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Poss egger, Horst Bischof; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24152-24161
Test-Time-Training (TTT) is an approach to cope with out-of-distribution (OOD) d ata by adapting a trained model to distribution shifts occurring at test-time. W e propose to perform this adaptation via Activation Matching (ActMAD): We analyz e activations of the model and align activation statistics of the OOD test data to those of the training data. In contrast to existing methods, which model the distribution of entire channels in the ultimate layer of the feature extractor, we model the distribution of each feature in multiple layers across the network. This results in a more fine-grained supervision and makes ActMAD attain state o f the art performance on CIFAR-100C and Imagenet-C. ActMAD is also architecture- and task-agnostic, which lets us go beyond image classification, and score 15.4 % improvement over previous approaches when evaluating a KITTI-trained object de tector on KITTI-Fog. Our experiments highlight that ActMAD can be applied to onl ine adaptation in realistic scenarios, requiring little data to attain its full performance.
****
DKM: Dense Kernelized Feature Matching for Geometry Estimation
Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, Michael Felsberg; Proceed ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17765-17775
Feature matching is a challenging computer vision task that involves finding cor respondences between two images of a 3D scene. In this paper we consider the den se approach instead of the more common sparse paradigm, thus striving to find al l correspondences. Perhaps counter-intuitively, dense methods have previously sh own inferior performance to their sparse and semi-sparse counterparts for estima tion of two-view geometry. This changes with our novel dense method, which outpe rforms both dense and sparse methods on geometry estimation. The novelty is thre efold: First, we propose a kernel regression global matcher. Secondly, we propos e warp refinement through stacked feature maps and depthwise convolution kernels . Thirdly, we propose learning dense confidence through consistent depth and a b alanced sampling approach for dense confidence maps. Through extensive experimen ts we confirm that our proposed dense method, Dense Kernelized Feature Matching, sets a new state-of-the-art on multiple geometry estimation benchmarks. In part icular, we achieve an improvement on MegaDepth-1500 of +4.9 and +8.9 AUC@5 compa red to the best previous sparse method and dense method respectively. Our code i s provided at the following repository: https://github.com/Parskatt/DKM
****
Image Cropping With Spatial-Aware Feature and Rank Consistency
Chao Wang, Li Niu, Bo Zhang, Liqing Zhang; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10052-10061
Image cropping aims to find visually appealing crops in an image. Despite the gr eat progress made by previous methods, they are weak in capturing the spatial re lationship between crops and aesthetic elements (e.g., salient objects, semantic edges). Besides, due to the high annotation cost of labeled data, the potential of unlabeled data awaits to be excavated. To address the first issue, we propos

e spatial-aware feature to encode the spatial relationship between candidate cro
ps and aesthetic elements, by feeding the concatenation of crop mask and selecti
vely aggregated feature maps to a light-weighted encoder. To address the second
issue, we train a pair-wise ranking classifier on labeled images and transfer su
ch knowledge to unlabeled images to enforce rank consistency. Experimental resul
ts on the benchmark datasets show that our proposed method performs favorably ag
ainst state-of-the-art methods.
****
SVGformer: Representation Learning for Continuous Vector Graphics Using Transfor
mers
Defu Cao, Zhaowen Wang, Jose Echevarria, Yan Liu; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10093-1010
2
Advances in representation learning have led to great success in understanding a
nd generating data in various domains. However, in modeling vector graphics data
, the pure data-driven approach often yields unsatisfactory results in downstrea
m tasks as existing deep learning methods often require the quantization of SVG
parameters and cannot exploit the geometric properties explicitly. In this paper
, we propose a transformer-based representation learning model (SVGformer) that
directly operates on continuous input values and manipulates the geometric infor
mation of SVG to encode outline details and long-distance dependencies. SVGfomer
 can be used for various downstream tasks: reconstruction, classification, inter
polation, retrieval, etc. We have conducted extensive experiments on vector font
 and icon datasets to show that our model can capture high-quality representatio
n information and outperform the previous state-of-the-art on downstream tasks s
ignificantly.
****
Structured 3D Features for Reconstructing Controllable Avatars
Enric Corona, Mihai Zanfir, Thiemo Alldieck, Eduard Gabriel Bazavan, Andrei Zanf
ir, Cristian Sminchisescu; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2023, pp. 16954-16964
We introduce Structured 3D Features, a model based on a novel implicit 3D repres
entation that pools pixel-aligned image features onto dense 3D points sampled fr
om a parametric, statistical human mesh surface. The 3D points have associated s
emantics and can move freely in 3D space. This allows for optimal coverage of th
e person of interest, beyond just the body shape, which in turn, additionally he
lps modeling accessories, hair, and loose clothing. Owing to this, we present a
complete 3D transformer-based attention framework which, given a single image of
 a person in an unconstrained pose, generates an animatable 3D reconstruction wi
th albedo and illumination decomposition, as a result of a single end-to-end mod
el, trained semi-supervised, and with no additional postprocessing. We show that
 our S3F model surpasses the previous state-of-the-art on various tasks, includi
ng monocular 3D reconstruction, as well as albedo & shading estimation. Moreover
, we show that the proposed methodology allows novel view synthesis, relighting,
 and re-posing the reconstruction, and can naturally be extended to handle multi
ple input images (e.g. different views of a person, or the same view, in differe
nt poses, in video). Finally, we demonstrate the editing capabilities of our mod
el for 3D virtual try-on applications.
****
Mask-Guided Matting in the Wild
Kwanyong Park, Sanghyun Woo, Seoung Wug Oh, In So Kweon, Joon-Young Lee; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2023, pp. 1992-2001
Mask-guided matting has shown great practicality compared to traditional trimap-
based methods. The mask-guided approach takes an easily-obtainable coarse mask a
s guidance and produces an accurate alpha matte. To extend the success toward pr
actical usage, we tackle mask-guided matting in the wild, which covers a wide ra
nge of categories in their complex context robustly. To this end, we propose a s
imple yet effective learning framework based on two core insights: 1) learning a
 generalized matting model that can better understand the given mask guidance an

d 2) leveraging weak supervision datasets (e.g., instance segmentation dataset) to alleviate the limited diversity and scale of existing matting datasets. Extensive experimental results on multiple benchmarks, consisting of a newly proposed synthetic benchmark (Composition-Wild) and existing natural datasets, demonstrate the superiority of the proposed method. Moreover, we provide appealing results on new practical applications (e.g., panoptic matting and mask-guided video matting), showing the great generality and potential of our model.
****
Dynamic Conceptional Contrastive Learning for Generalized Category Discovery
Nan Pu, Zhun Zhong, Nicu Sebe; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7579-7588
Generalized category discovery (GCD) is a recently proposed open-world problem, which aims to automatically cluster partially labeled data. The main challenge is that the unlabeled data contain instances that are not only from known categories of the labeled data but also from novel categories. This leads traditional novel category discovery (NCD) methods to be incapacitated for GCD, due to their assumption of unlabeled data are only from novel categories. One effective way for GCD is applying self-supervised learning to learn discriminate representation for unlabeled data. However, this manner largely ignores underlying relationships between instances of the same concepts (e.g., class, super-class, and sub-class), which results in inferior representation learning. In this paper, we propose a Dynamic Conceptional Contrastive Learning (DCCL) framework, which can effectively improve clustering accuracy by alternately estimating underlying visual conceptions and learning conceptional representation. In addition, we design a dynamic conception generation and update mechanism, which is able to ensure consistent conception learning and thus further facilitate the optimization of DCCL. Extensive experiments show that DCCL achieves new state-of-the-art performances on six generic and fine-grained visual recognition datasets, especially on fine-grained ones. For example, our method significantly surpasses the best competitor by 16.2% on the new classes for the CUB-200 dataset. Code is available at https://github.com/TPCD/DCCL
****
Neumann Network With Recursive Kernels for Single Image Defocus Deblurring
Yuhui Quan, Zicong Wu, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5754-5763
Single image defocus deblurring (SIDD) refers to recovering an all-in-focus image from a defocused blurry one. It is a challenging recovery task due to the spatially-varying defocus blurring effects with significant size variation. Motivated by the strong correlation among defocus kernels of different sizes and the blob-type structure of defocus kernels, we propose a learnable recursive kernel representation (RKR) for defocus kernels that expresses a defocus kernel by a linear combination of recursive, separable and positive atom kernels, leading to a compact yet effective and physics-encoded parametrization of the spatially-varying defocus blurring process. Afterwards, a physics-driven and efficient deep model with a cross-scale fusion structure is presented for SIDD, with inspirations from the truncated Neumann series for approximating the matrix inversion of the RKR-based blurring operator. In addition, a reblurring loss is proposed to regularize the RKR learning. Extensive experiments show that, our proposed approach significantly outperforms existing ones, with a model size comparable to that of the top methods.
****
Active Finetuning: Exploiting Annotation Budget in the Pretraining-Finetuning Paradigm
Yichen Xie, Han Lu, Junchi Yan, Xiaokang Yang, Masayoshi Tomizuka, Wei Zhan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23715-23724
Given the large-scale data and the high annotation cost, pretraining-finetuning becomes a popular paradigm in multiple computer vision tasks. Previous research has covered both the unsupervised pretraining and supervised finetuning in this paradigm, while little attention is paid to exploiting the annotation budget for

finetuning. To fill in this gap, we formally define this new active finetuning task focusing on the selection of samples for annotation in the pretraining-fine tuning paradigm. We propose a novel method called ActiveFT for active finetuning task to select a subset of data distributing similarly with the entire unlabeled pool and maintaining enough diversity by optimizing a parametric model in the continuous space. We prove that the Earth Mover's distance between the distributions of the selected subset and the entire data pool is also reduced in this process. Extensive experiments show the leading performance and high efficiency of ActiveFT superior to baselines on both image classification and semantic segmentation.
****
Learning Attribute and Class-Specific Representation Duet for Fine-Grained Fashion Analysis
Yang Jiao, Yan Gao, Jingjing Meng, Jin Shang, Yi Sun; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11050-11059
Fashion representation learning involves the analysis and understanding of various visual elements at different granularities and the interactions among them. Existing works often learn fine-grained fashion representations at the attribute-level without considering their relationships and inter-dependencies across different classes. In this work, we propose to learn an attribute and class specific fashion representation duet to better model such attribute relationships and inter-dependencies by leveraging prior knowledge about the taxonomy of fashion attributes and classes. Through two sub-networks for the attributes and classes, respectively, our proposed an embedding network progressively learn and refine the visual representation of a fashion image to improve its robustness for fashion retrieval. A multi-granularity loss consisting of attribute-level and class-level losses is proposed to introduce appropriate inductive bias to learn across different granularities of the fashion representations. Experimental results on three benchmark datasets demonstrate the effectiveness of our method, which outperforms the state-of-the-art methods with a large margin.
****
Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection
Yi Wang, Ruili Wang, Xin Fan, Tianzhu Wang, Xiangjian He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10031-10040
Salient object detection (SOD) aims to mimic the human visual system (HVS) and cognition mechanisms to identify and segment salient objects. However, due to the complexity of these mechanisms, current methods are not perfect. Accuracy and robustness need to be further improved, particularly in complex scenes with multiple objects and background clutter. To address this issue, we propose a novel approach called Multiple Enhancement Network (MENet) that adopts the boundary sensibility, content integrity, iterative refinement, and frequency decomposition mechanisms of HVS. A multi-level hybrid loss is firstly designed to guide the network to learn pixel-level, region-level, and object-level features. A flexible multiscale feature enhancement module (ME-Module) is then designed to gradually aggregate and refine global or detailed features by changing the size order of the input feature sequence. An iterative training strategy is used to enhance boundary features and adaptive features in the dual-branch decoder of MENet. Comprehensive evaluations on six challenging benchmark datasets show that MENet achieves state-of-the-art results. Both the codes and results are publicly available at https://github.com/yiwangtz/MENet.
****
Leveraging Temporal Context in Low Representational Power Regimes
Camilo L. Fosco, SouYoung Jin, Emilie Josephs, Aude Oliva; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10693-10703
Computer vision models are excellent at identifying and exploiting regularities in the world. However, it is computationally costly to learn these regularities from scratch. This presents a challenge for low-parameter models, like those run

ning on edge devices (e.g. smartphones). Can the performance of models with low representational power be improved by supplementing training with additional information about these statistical regularities? We explore this in the domains of action recognition and action anticipation, leveraging the fact that actions are typically embedded in stereotypical sequences. We introduce the Event Transition Matrix (ETM), computed from action labels in an untrimmed video dataset, which captures the temporal context of a given action, operationalized as the likelihood that it was preceded or followed by each other action in the set. We show that including information from the ETM during training improves action recognition and anticipation performance on various egocentric video datasets. Through ablation and control studies, we show that the coherent sequence of information captured by our ETM is key to this effect, and we find that the benefit of this explicit representation of temporal context is most pronounced for smaller models. Code, matrices and models are available in our project page: https://camilofosco.com/etm_website.
****
Guided Recommendation for Model Fine-Tuning
Hao Li, Charless Fowlkes, Hao Yang, Onkar Dabeer, Zhuowen Tu, Stefano Soatto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3633-3642
Model selection is essential for reducing the search cost of the best pre-trained model over a large-scale model zoo for a downstream task. After analyzing recent hand-designed model selection criteria with 400+ ImageNet pre-trained models and 40 downstream tasks, we find that they can fail due to invalid assumptions and intrinsic limitations. The prior knowledge on model capacity and dataset also can not be easily integrated into the existing criteria. To address these issues, we propose to convert model selection as a recommendation problem and to learn from the past training history. Specifically, we characterize the meta information of datasets and models as features, and use their transfer learning performance as the guided score. With thousands of historical training jobs, a recommendation system can be learned to predict the model selection score given the features of the dataset and the model as input. Our approach enables integrating existing model selection scores as additional features and scales with more historical data. We evaluate the prediction accuracy with 22 pre-trained models over 40 downstream tasks. With extensive evaluations, we show that the learned approach can outperform prior hand-designed model selection methods significantly when relevant training history is available.
****
Masked Image Training for Generalizable Deep Image Denoising
Haoyu Chen, Jinjin Gu, Yihao Liu, Salma Abdel Magid, Chao Dong, Qiong Wang, Hanspeter Pfister, Lei Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1692-1703
When capturing and storing images, devices inevitably introduce noise. Reducing this noise is a critical task called image denoising. Deep learning has become the de facto method for image denoising, especially with the emergence of Transformer-based models that have achieved notable state-of-the-art results on various image tasks. However, deep learning-based methods often suffer from a lack of generalization ability. For example, deep models trained on Gaussian noise may perform poorly when tested on other noise distributions. To address this issue, we present a novel approach to enhance the generalization performance of denoising networks, known as masked training. Our method involves masking random pixels of the input image and reconstructing the missing information during training. We also mask out the features in the self-attention layers to avoid the impact of training-testing inconsistency. Our approach exhibits better generalization ability than other deep learning models and is directly applicable to real-world scenarios. Additionally, our interpretability analysis demonstrates the superiority of our method.
****
In-Hand 3D Object Scanning From an RGB Sequence
Shreyas Hampali, Tomas Hodan, Luan Tran, Lingni Ma, Cem Keskin, Vincent Lepetit;

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti on (CVPR), 2023, pp. 17079-17088
We propose a method for in-hand 3D scanning of an unknown object with a monocula r camera. Our method relies on a neural implicit surface representation that cap tures both the geometry and the appearance of the object, however, by contrast w ith most NeRF-based methods, we do not assume that the camera-object relative po ses are known. Instead, we simultaneously optimize both the object shape and the pose trajectory. As direct optimization over all shape and pose parameters is p rone to fail without coarse-level initialization, we propose an incremental appr oach that starts by splitting the sequence into carefully selected overlapping s egments within which the optimization is likely to succeed. We reconstruct the o bject shape and track its poses independently within each segment, then merge al l the segments before performing a global optimization. We show that our method is able to reconstruct the shape and color of both textured and challenging text ure-less objects, outperforms classical methods that rely only on appearance fea tures, and that its performance is close to recent methods that assume known cam era poses.
****
Zero-Shot Referring Image Segmentation With Global-Local Context Features
Seonghoon Yu, Paul Hongsuck Seo, Jeany Son; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19456-19465
Referring image segmentation (RIS) aims to find a segmentation mask given a refe rring expression grounded to a region of the input image. Collecting labelled da tasets for this task, however, is notoriously costly and labor-intensive. To ove rcome this issue, we propose a simple yet effective zero-shot referring image se gmentation method by leveraging the pre-trained cross-modal knowledge from CLIP. In order to obtain segmentation masks grounded to the input text, we propose a mask-guided visual encoder that captures global and local contextual information of an input image. By utilizing instance masks obtained from off-the-shelf mask proposal techniques, our method is able to segment fine-detailed instance-level groundings. We also introduce a global-local text encoder where the global feat ure captures complex sentence-level semantics of the entire input expression whi le the local feature focuses on the target noun phrase extracted by a dependency parser. In our experiments, the proposed method outperforms several zero-shot b aselines of the task and even the weakly supervised referring expression segment ation method with substantial margins. Our code is available at https://github.c om/Seonghoon-Yu/Zero-shot-RIS.
****
SketchXAI: A First Look at Explainability for Human Sketches
Zhiyu Qu, Yulia Gryaditskaya, Ke Li, Kaiyue Pang, Tao Xiang, Yi-Zhe Song; Procee dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP R), 2023, pp. 23327-23337
This paper, for the very first time, introduces human sketches to the landscape of XAI (Explainable Artificial Intelligence). We argue that sketch as a "human-c entred" data form, represents a natural interface to study explainability. We fo cus on cultivating sketch-specific explainability designs. This starts by identi fying strokes as a unique building block that offers a degree of flexibility in object construction and manipulation impossible in photos. Following this, we de sign a simple explainability-friendly sketch encoder that accommodates the intri nsic properties of strokes: shape, location, and order. We then move on to defin e the first ever XAI task for sketch, that of stroke location inversion SLI. Jus t as we have heat maps for photos, and correlation matrices for text, SLI offers an explainability angle to sketch in terms of asking a network how well it can recover stroke locations of an unseen sketch. We offer qualitative results for r eaders to interpret as snapshots of the SLI process in the paper, and as GIFs on the project page. A minor but interesting note is that thanks to its sketch-spe cific design, our sketch encoder also yields the best sketch recognition accurac y to date while having the smallest number of parameters. The code is available at https://sketchxai.github.io.
****

Omni3D: A Large Benchmark and Model for 3D Object Detection in the Wild
Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, Georgia Gkioxari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13154-13164
Recognizing scenes and objects in 3D from a single image is a longstanding goal of computer vision with applications in robotics and AR/VR. For 2D recognition, large datasets and scalable solutions have led to unprecedented advances. In 3D, existing benchmarks are small in size and approaches specialize in few object categories and specific domains, e.g. urban driving scenes. Motivated by the success of 2D recognition, we revisit the task of 3D object detection by introducing a large benchmark, called Omni3D. Omni3D re-purposes and combines existing datasets resulting in 234k images annotated with more than 3 million instances and 98 categories. 3D detection at such scale is challenging due to variations in camera intrinsics and the rich diversity of scene and object types. We propose a model, called Cube R-CNN, designed to generalize across camera and scene types with a unified approach. We show that Cube R-CNN outperforms prior works on the larger Omni3D and existing benchmarks. Finally, we prove that Omni3D is a powerful dataset for 3D object recognition and show that it improves single-dataset performance and can accelerate learning on new smaller datasets via pre-training.
****
OT-Filter: An Optimal Transport Filter for Learning With Noisy Labels
Chuanwen Feng, Yilong Ren, Xike Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16164-16174
The success of deep learning is largely attributed to the training over clean data. However, data is often coupled with noisy labels in practice. Learning with noisy labels is challenging because the performance of the deep neural networks (DNN) drastically degenerates, due to confirmation bias caused by the network memorization over noisy labels. To alleviate that, a recent prominent direction is on sample selection, which retrieves clean data samples from noisy samples, so as to enhance the model's robustness and tolerance to noisy labels. In this paper, we revamp the sample selection from the perspective of optimal transport theory and propose a novel method, called the OT-Filter. The OT-Filter provides geometrically meaningful distances and preserves distribution patterns to measure the data discrepancy, thus alleviating the confirmation bias. Extensive experiments on benchmarks, such as Clothing1M and ANIMAL-10N, show that the performance of the OT- Filter outperforms its counterparts. Meanwhile, results on benchmarks with synthetic labels, such as CIFAR-10/100, show the superiority of the OT-Filter in handling data labels of high noise.
****
Rebalancing Batch Normalization for Exemplar-Based Class-Incremental Learning
Sungmin Cha, Sungjun Cho, Dasol Hwang, Sunwon Hong, Moontae Lee, Taesup Moon; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20127-20136
Batch Normalization (BN) and its variants has been extensively studied for neural nets in various computer vision tasks, but relatively little work has been dedicated to studying the effect of BN in continual learning. To that end, we develop a new update patch for BN, particularly tailored for the exemplar-based class-incremental learning (CIL). The main issue of BN in CIL is the imbalance of training data between current and past tasks in a mini-batch, which makes the empirical mean and variance as well as the learnable affine transformation parameters of BN heavily biased toward the current task --- contributing to the forgetting of past tasks. While one of the recent BN variants has been developed for "online" CIL, in which the training is done with a single epoch, we show that their method does not necessarily bring gains for "offline" CIL, in which a model is trained with multiple epochs on the imbalanced training data. The main reason for the ineffectiveness of their method lies in not fully addressing the data imbalance issue, especially in computing the gradients for learning the affine transformation parameters of BN. Accordingly, our new hyperparameter-free variant, dubbed as Task-Balanced BN (TBBN), is proposed to more correctly resolve the imbalance issue by making a horizontally-concatenated task-balanced batch using both re

shape and repeat operations during training. Based on our experiments on class incremental learning of CIFAR-100, ImageNet-100, and five dissimilar task datasets, we demonstrate that our TBBN, which works exactly the same as the vanilla BN in the inference time, is easily applicable to most existing exemplar-based offline CIL algorithms and consistently outperforms other BN variants.
****

OmniVidar: Omnidirectional Depth Estimation From Multi-Fisheye Images
Sheng Xie, Daochuan Wang, Yun-Hui Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21529-21538
Estimating depth from four large field of view (FoV) cameras has been a difficult and understudied problem. In this paper, we proposed a novel and simple system that can convert this difficult problem into easier binocular depth estimation. We name this system OmniVidar, as its results are similar to LiDAR, but rely only on vision. OmniVidar contains three components: (1) a new camera model to address the shortcomings of existing models, (2) a new multi-fisheye camera based epipolar rectification method for solving the image distortion and simplifying the depth estimation problem, (3) an improved binocular depth estimation network, which achieves a better balance between accuracy and efficiency. Unlike other omnidirectional stereo vision methods, OmniVidar does not contain any 3D convolution, so it can achieve higher resolution depth estimation at fast speed. Results demonstrate that OmniVidar outperforms all other methods in terms of accuracy and performance.
****

RWSC-Fusion: Region-Wise Style-Controlled Fusion Network for the Prohibited X-Ray Security Image Synthesis
Luwen Duan, Min Wu, Lijian Mao, Jun Yin, Jianping Xiong, Xi Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22398-22407
Automatic prohibited item detection in security inspection X-ray images is necessary for transportation.The abundance and diversity of the X-ray security images with prohibited item, termed as prohibited X-ray security images, are essential for training the detection model. In order to solve the data insufficiency, we propose a RegionWise Style-Controlled Fusion (RWSC-Fusion) network, which superimposes the prohibited items onto the normal X-ray security images, to synthesize the prohibited X-ray security images. The proposed RWSC-Fusion innovates both network structure and loss functions to generate more realistic X-ray security images. Specifically, a RWSCFusion module is designed to enable the region-wise fusion by controlling the appearance of the overlapping region with novel modulation parameters. In addition, an EdgeAttention (EA) module is proposed to effectively improve the sharpness of the synthetic images. As for the unsupervised loss function, we propose the Luminance loss in Logarithmic form (LL) and Correlation loss of Saturation Difference (CSD), to optimize the fused X-ray security images in terms of luminance and saturation. We evaluate the authenticity and the training effect of the synthetic X-ray security images on private and public SIXray dataset. The results confirm that our synthetic images are reliable enough to augment the prohibited Xray security images.
****

Octree Guided Unoriented Surface Reconstruction
Chamin Hewa Koneputugodage, Yizhak Ben-Shabat, Stephen Gould; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16717-16726
We address the problem of surface reconstruction from unoriented point clouds. Implicit neural representations (INRs) have become popular for this task, but when information relating to the inside versus outside of a shape is not available (such as shape occupancy, signed distances or surface normal orientation) optimization relies on heuristics and regularizers to recover the surface. These methods can be slow to converge and easily get stuck in local minima. We propose a two-step approach, OG-INR, where we (1) construct a discrete octree and label what is inside and outside (2) optimize for a continuous and high-fidelity shape using an INR that is initially guided by the octree's labelling. To solve for our l

abelling, we propose an energy function over the discrete structure and provide an efficient move-making algorithm that explores many possible labellings. Furthermore we show that we can easily inject knowledge into the discrete octree, providing a simple way to influence the result from the continuous INR. We evaluate the effectiveness of our approach on two unoriented surface reconstruction datasets and show competitive performance compared to other unoriented, and some oriented, methods. Our results show that the exploration by the move-making algorithm avoids many of the bad local minima reached by purely gradient descent optimized methods (see Figure 1).
****
Rigidity-Aware Detection for 6D Object Pose Estimation
Yang Hai, Rui Song, Jiaojiao Li, Mathieu Salzmann, Yinlin Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8927-8936
Most recent 6D object pose estimation methods first use object detection to obtain 2D bounding boxes before actually regressing the pose. However, the general object detection methods they use are ill-suited to handle cluttered scenes, thus producing poor initialization to the subsequent pose network. To address this, we propose a rigidity-aware detection method exploiting the fact that, in 6D pose estimation, the target objects are rigid. This lets us introduce an approach to sampling positive object regions from the entire visible object area during training, instead of naively drawing samples from the bounding box center where the object might be occluded. As such, every visible object part can contribute to the final bounding box prediction, yielding better detection robustness. Key to the success of our approach is a visibility map, which we propose to build using a minimum barrier distance between every pixel in the bounding box and the box boundary. Our results on seven challenging 6D pose estimation datasets evidence that our method outperforms general detection frameworks by a large margin. Furthermore, combined with a pose regression network, we obtain state-of-the-art pose estimation results on the challenging BOP benchmark.
****
ToThePoint: Efficient Contrastive Learning of 3D Point Clouds via Recycling
Xinglin Li, Jiajing Chen, Jinhui Ouyang, Hanhui Deng, Senem Velipasalar, Di Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21781-21790
Recent years have witnessed significant developments in point cloud processing, including classification and segmentation. However, supervised learning approaches need a lot of well-labeled data for training, and annotation is labor- and time-intensive. Self-supervised learning, on the other hand, uses unlabeled data, and pre-trains a backbone with a pretext task to extract latent representations to be used with the downstream tasks. Compared to 2D images, self-supervised learning of 3D point clouds is under-explored. Existing models, for self-supervised learning of 3D point clouds, rely on a large number of data samples, and require significant amount of computational resources and training time. To address this issue, we propose a novel contrastive learning approach, referred to as ToThePoint. Different from traditional contrastive learning methods, which maximize agreement between features obtained from a pair of point clouds formed only with different types of augmentation, ToThePoint also maximizes the agreement between the permutation invariant features and features discarded after max pooling. We first perform self-supervised learning on the ShapeNet dataset, and then evaluate the performance of the network on different downstream tasks. In the downstream task experiments, performed on the ModelNet40, ModelNet40C, ScanobjectNN and ShapeNet-Part datasets, our proposed ToThePoint achieves competitive, if not better results compared to the state-of-the-art baselines, and does so with significantly less training time (200 times faster than baselines)
****
Clover: Towards a Unified Video-Language Alignment and Fusion Model
Jingjia Huang, Yinan Li, Jiashi Feng, Xinglong Wu, Xiaoshuai Sun, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14856-14866

Building a universal video-language model for solving various video understandin
g tasks (e.g., text-video retrieval, video question answering) is an open challe
nge to the machine learning field. Towards this goal, most recent works build th
e model by stacking uni-modal and cross-modal feature encoders and train it with
 pair-wise contrastive pre-text tasks. Though offering attractive generality, th
e resulted models have to compromise between efficiency and performance. They mo
stly adopt different architectures to deal with different downstream tasks. We f
ind this is because the pair-wise training cannot well align and fuse features f
rom different modalities. We then introduce Clover--a Correlated Video-Language
pre-training method--towards a universal video-language model for solving multip
le video understanding tasks with neither performance nor efficiency compromise.
 It improves cross-modal feature alignment and fusion via a novel tri-modal alig
nment pre-training task. Additionally, we propose to enhance the tri-modal align
ment via incorporating learning from semantic masked samples and a new pair-wise
 ranking loss. Clover establishes new state-of-the-arts on multiple downstream t
asks, including three retrieval tasks for both zero-shot and fine-tuning setting
s, and eight video question answering tasks. Codes and pre-trained models will b
e released at https://github.com/LeeYN-43/Clover.
****
Weakly Supervised Monocular 3D Object Detection Using Multi-View Projection and
Direction Consistency
Runzhou Tao, Wencheng Han, Zhongying Qiu, Cheng-Zhong Xu, Jianbing Shen; Proceed
ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR
), 2023, pp. 17482-17492
Monocular 3D object detection has become a mainstream approach in automatic driv
ing for its easy application. A prominent advantage is that it does not need LiD
AR point clouds during the inference. However, most current methods still rely o
n 3D point cloud data for labeling the ground truths used in the training phase.
 This inconsistency between the training and inference makes it hard to utilize
the large-scale feedback data and increases the data collection expenses. To bri
dge this gap, we propose a new weakly supervised monocular 3D objection detectio
n method, which can train the model with only 2D labels marked on images. To be
specific, we explore three types of consistency in this task, i.e. the projectio
n, multi-view and direction consistency, and design a weakly-supervised architec
ture based on these consistencies. Moreover, we propose a new 2D direction label
ing method in this task to guide the model for accurate rotation direction predi
ction. Experiments show that our weakly-supervised method achieves comparable pe
rformance with some fully supervised methods. When used as a pre-training method
, our model can significantly outperform the corresponding fully-supervised base
line with only 1/3 3D labels.
****
Self-Supervised Learning From Images With a Joint-Embedding Predictive Architect
ure
Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Mi
chael Rabbat, Yann LeCun, Nicolas Ballas; Proceedings of the IEEE/CVF Conference
 on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15619-15629
This paper demonstrates an approach for learning highly semantic image represent
ations without relying on hand-crafted data-augmentations. We introduce the Imag
e-based Joint-Embedding Predictive Architecture (I-JEPA), a non-generative appro
ach for self-supervised learning from images. The idea behind I-JEPA is simple:
from a single context block, predict the representations of various target block
s in the same image. A core design choice to guide I-JEPA towards producing sema
ntic representations is the masking strategy; specifically, it is crucial to (a)
 sample target blocks with sufficiently large scale (semantic), and to (b) use a
 sufficiently informative (spatially distributed) context block. Empirically, wh
en combined with Vision Transformers, we find I-JEPA to be highly scalable. For
instance, we train a ViT-Huge/14 on ImageNet using 16 A100 GPUs in under 72 hour
s to achieve strong downstream performance across a wide range of tasks, from li
near classification to object counting and depth prediction.
****

EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding
Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, Jian Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19231-19242
3D visual grounding aims to find the object within point clouds mentioned by free-form natural language descriptions with rich semantic cues. However, existing methods either extract the sentence-level features coupling all words or focus more on object names, which would lose the word-level information or neglect other attributes. To alleviate these issues, we present EDA that Explicitly Decouples the textual attributes in a sentence and conducts Dense Alignment between such fine-grained language and point cloud objects. Specifically, we first propose a text decoupling module to produce textual features for every semantic component. Then, we design two losses to supervise the dense matching between two modalities: position alignment loss and semantic alignment loss. On top of that, we further introduce a new visual grounding task, locating objects without object names, which can thoroughly evaluate the model's dense alignment capacity. Through experiments, we achieve state-of-the-art performance on two widely-adopted 3D visual grounding datasets, ScanRefer and SR3D/NR3D, and obtain absolute leadership on our newly-proposed task. The source code is available at https://github.com/yanmin-wu/EDA.
****
A2J-Transformer: Anchor-to-Joint Transformer Network for 3D Interacting Hand Pose Estimation From a Single RGB Image
Changlong Jiang, Yang Xiao, Cunlin Wu, Mingyang Zhang, Jinghong Zheng, Zhiguo Cao, Joey Tianyi Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8846-8855
3D interacting hand pose estimation from a single RGB image is a challenging task, due to serious self-occlusion and inter-occlusion towards hands, confusing similar appearance patterns between 2 hands, ill-posed joint position mapping from 2D to 3D, etc.. To address these, we propose to extend A2J-the state-of-the-art depth-based 3D single hand pose estimation method-to RGB domain under interacting hand condition. Our key idea is to equip A2J with strong local-global aware ability to well capture interacting hands' local fine details and global articulated clues among joints jointly. To this end, A2J is evolved under Transformer's non-local encoding-decoding framework to build A2J-Transformer. It holds 3 main advantages over A2J. First, self-attention across local anchor points is built to make them global spatial context aware to better capture joints' articulation clues for resisting occlusion. Secondly, each anchor point is regarded as learnable query with adaptive feature learning for facilitating pattern fitting capacity, instead of having the same local representation with the others. Last but not least, anchor point locates in 3D space instead of 2D as in A2J, to leverage 3D pose prediction. Experiments on challenging InterHand 2.6M demonstrate that, A2J-Transformer can achieve state-of-the-art model-free performance (3.38mm MPJPE advancement in 2-hand case) and can also be applied to depth domain with strong generalization.
****
The Treasure Beneath Multiple Annotations: An Uncertainty-Aware Edge Detector
Caixia Zhou, Yaping Huang, Mengyang Pu, Qingji Guan, Li Huang, Haibin Ling; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15507-15517
Deep learning-based edge detectors heavily rely on pixel-wise labels which are often provided by multiple annotators. Existing methods fuse multiple annotations using a simple voting process, ignoring the inherent ambiguity of edges and labeling bias of annotators. In this paper, we propose a novel uncertainty-aware edge detector (UAED), which employs uncertainty to investigate the subjectivity and ambiguity of diverse annotations. Specifically, we first convert the deterministic label space into a learnable Gaussian distribution, whose variance measures the degree of ambiguity among different annotations. Then we regard the learned variance as the estimated uncertainty of the predicted edge maps, and pixels with higher uncertainty are likely to be hard samples for edge detection. Therefor

e we design an adaptive weighting loss to emphasize the learning from those pixels with high uncertainty, which helps the network to gradually concentrate on the important pixels. UAED can be combined with various encoder-decoder backbones, and the extensive experiments demonstrate that UAED achieves superior performance consistently across multiple edge detection benchmarks. The source code is available at https://github.com/ZhouCX117/UAED.
****
DP-NeRF: Deblurred Neural Radiance Field With Physical Scene Priors
Dogyoon Lee, Minhyeok Lee, Chajin Shin, Sangyoun Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12386-12396
Neural Radiance Field (NeRF) has exhibited outstanding three-dimensional (3D) reconstruction quality via the novel view synthesis from multi-view images and paired calibrated camera parameters. However, previous NeRF-based systems have been demonstrated under strictly controlled settings, with little attention paid to less ideal scenarios, including with the presence of noise such as exposure, illumination changes, and blur. In particular, though blur frequently occurs in real situations, NeRF that can handle blurred images has received little attention. The few studies that have investigated NeRF for blurred images have not considered geometric and appearance consistency in 3D space, which is one of the most important factors in 3D reconstruction. This leads to inconsistency and the degradation of the perceptual quality of the constructed scene. Hence, this paper proposes a DP-NeRF, a novel clean NeRF framework for blurred images, which is constrained with two physical priors. These priors are derived from the actual blurring process during image acquisition by the camera. DP-NeRF proposes rigid blurring kernel to impose 3D consistency utilizing the physical priors and adaptive weight proposal to refine the color composition error in consideration of the relationship between depth and blur. We present extensive experimental results for synthetic and real scenes with two types of blur: camera motion blur and defocus blur. The results demonstrate that DP-NeRF successfully improves the perceptual quality of the constructed NeRF ensuring 3D geometric and appearance consistency. We further demonstrate the effectiveness of our model with comprehensive ablation analysis.
****
MixPHM: Redundancy-Aware Parameter-Efficient Tuning for Low-Resource Visual Question Answering
Jingjing Jiang, Nanning Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24203-24213
Recently, finetuning pretrained vision-language models (VLMs) has been a prevailing paradigm for achieving state-of-the-art performance in VQA. However, as VLMs scale, it becomes computationally expensive, storage inefficient, and prone to overfitting when tuning full model parameters for a specific task in low-resource settings. Although current parameter-efficient tuning methods dramatically reduce the number of tunable parameters, there still exists a significant performance gap with full finetuning. In this paper, we propose MixPHM, a redundancy-aware parameter-efficient tuning method that outperforms full finetuning in low-resource VQA. Specifically, MixPHM is a lightweight module implemented by multiple PHM-experts in a mixture-of-experts manner. To reduce parameter redundancy, we reparameterize expert weights in a low-rank subspace and share part of the weights inside and across MixPHM. Moreover, based on our quantitative analysis of representation redundancy, we propose Redundancy Regularization, which facilitates MixPHM to reduce task-irrelevant redundancy while promoting task-relevant correlation. Experiments conducted on VQA v2, GQA, and OK-VQA with different low-resource settings show that our MixPHM outperforms state-of-the-art parameter-efficient methods and is the only one consistently surpassing full finetuning.
****
Self-Supervised Blind Motion Deblurring With Deep Expectation Maximization
Ji Li, Weixi Wang, Yuesong Nan, Hui Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13986-13996
When taking a picture, any camera shake during the shutter time can result in a

blurred image. Recovering a sharp image from the one blurred by camera shake is a challenging yet important problem. Most existing deep learning methods use sup ervised learning to train a deep neural network (DNN) on a dataset of many pairs of blurred/latent images. In contrast, this paper presents a dataset-free deep learning method for removing uniform and non-uniform blur effects from images of static scenes. Our method involves a DNN-based re-parametrization of the latent image, and we propose a Monte Carlo Expectation Maximization (MCEM) approach to train the DNN without requiring any latent images. The Monte Carlo simulation i s implemented via Langevin dynamics. Experiments showed that the proposed method outperforms existing methods significantly in removing motion blur from images of static scenes.
****
DeAR: Debiasing Vision-Language Models With Additive Residuals
Ashish Seth, Mayur Hemani, Chirag Agarwal; Proceedings of the IEEE/CVF Conferenc e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6820-6829
Large pre-trained vision-language models (VLMs) reduce the time for developing p redictive models for various vision-grounded language downstream tasks by provid ing rich, adaptable image and text representations. However, these models suffer from societal biases owing to the skewed distribution of various identity group s in the training data. These biases manifest as the skewed similarity between t he representations for specific text concepts and images of people of different identity groups and, therefore, limit the usefulness of such models in real-worl d high-stakes applications. In this work, we present DeAR (Debiasing with Additi ve Residuals), a novel debiasing method that learns additive residual image repr esentations to offset the original representations, ensuring fair output represe ntations. In doing so, it reduces the ability of the representations to distingu ish between the different identity groups. Further, we observe that the current fairness tests are performed on limited face image datasets that fail to indicat e why a specific text concept should/should not apply to them. To bridge this ga p and better evaluate DeAR, we introduce a new context-based bias benchmarking d ataset - the Protected Attribute Tag Association (PATA) dataset for evaluating t he fairness of large pre-trained VLMs. Additionally, PATA provides visual contex t for a diverse human population in different scenarios with both positive and n egative connotations. Experimental results for fairness and zero-shot performanc e preservation using multiple datasets demonstrate the efficacy of our framework .
****
E2PN: Efficient SE(3)-Equivariant Point Network
Minghan Zhu, Maani Ghaffari, William A. Clark, Huei Peng; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12 23-1232
This paper proposes a convolution structure for learning SE(3)-equivariant featu res from 3D point clouds. It can be viewed as an equivariant version of kernel p oint convolutions (KPConv), a widely used convolution form to process point clou d data. Compared with existing equivariant networks, our design is simple, light weight, fast, and easy to be integrated with existing task-specific point cloud learning pipelines. We achieve these desirable properties by combining group con volutions and quotient representations. Specifically, we discretize SO(3) to fin ite groups for their simplicity while using SO(2) as the stabilizer subgroup to form spherical quotient feature fields to save computations. We also propose a p ermutation layer to recover SO(3) features from spherical features to preserve t he capacity to distinguish rotations. Experiments show that our method achieves comparable or superior performance in various tasks, including object classifica tion, pose estimation, and keypoint-matching, while consuming much less memory a nd running faster than existing work. The proposed method can foster the develop ment of equivariant models for real-world applications based on point clouds.
****
Understanding Masked Image Modeling via Learning Occlusion Invariant Feature
Xiangwen Kong, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6241-6251

Recently, Masked Image Modeling (MIM) achieves great success in self-supervised visual recognition. However, as a reconstruction-based framework, it is still an open question to understand how MIM works, since MIM appears very different from previous well-studied siamese approaches such as contrastive learning. In this paper, we propose a new viewpoint: MIM implicitly learns occlusion-invariant features, which is analogous to other siamese methods while the latter learns other invariance. By relaxing MIM formulation into an equivalent siamese form, MIM methods can be interpreted in a unified framework with conventional methods, among which only a) data transformations, i.e. what invariance to learn, and b) similarity measurements are different. Furthermore, taking MAE (He et al., 2021) as a representative example of MIM, we empirically find the success of MIM models relates a little to the choice of similarity functions, but the learned occlusion invariant feature introduced by masked image -- it turns out to be a favored initialization for vision transformers, even though the learned feature could be less semantic. We hope our findings could inspire researchers to develop more powerful self-supervised methods in computer vision community.
****
Grounding Counterfactual Explanation of Image Classifiers to Textual Concept Space
Siwon Kim, Jinoh Oh, Sungjin Lee, Seunghak Yu, Jaeyoung Do, Tara Taghavi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10942-10950
Concept-based explanation aims to provide concise and human-understandable explanations of an image classifier. However, existing concept-based explanation methods typically require a significant amount of manually collected concept-annotated images. This is costly and runs the risk of human biases being involved in the explanation. In this paper, we propose counterfactual explanation with text-driven concepts (CounTEX), where the concepts are defined only from text by leveraging a pre-trained multi-modal joint embedding space without additional concept-annotated datasets. A conceptual counterfactual explanation is generated with text-driven concepts. To utilize the text-driven concepts defined in the joint embedding space to interpret target classifier outcome, we present a novel projection scheme for mapping the two spaces with a simple yet effective implementation. We show that CounTEX generates faithful explanations that provide a semantic understanding of model decision rationale robust to human bias.
****
A Dynamic Multi-Scale Voxel Flow Network for Video Prediction
Xiaotao Hu, Zhewei Huang, Ailin Huang, Jun Xu, Shuchang Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6121-6131
The performance of video prediction has been greatly boosted by advanced deep neural networks. However, most of the current methods suffer from large model sizes and require extra inputs, e.g., semantic/depth maps, for promising performance. For efficiency consideration, in this paper, we propose a Dynamic Multi-scale Voxel Flow Network (DMVFN) to achieve better video prediction performance at lower computational costs with only RGB images, than previous methods. The core of our DMVFN is a differentiable routing module that can effectively perceive the motion scales of video frames. Once trained, our DMVFN selects adaptive sub-networks for different inputs at the inference stage. Experiments on several benchmarks demonstrate that our DMVFN is an order of magnitude faster than Deep Voxel Flow and surpasses the state-of-the-art iterative-based OPT on generated image quality. Our code and demo are available at https://huxiaotaostasy.github.io/DMVFN/.
****
UniDistill: A Universal Cross-Modality Knowledge Distillation Framework for 3D Object Detection in Bird's-Eye View
Shengchao Zhou, Weizhou Liu, Chen Hu, Shuchang Zhou, Chao Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5116-5125
In the field of 3D object detection for autonomous driving, the sensor portfolio

 including multi-modality and single-modality is diverse and complex. Since the multi-modal methods have system complexity while the accuracy of single-modal ones is relatively low, how to make a tradeoff between them is difficult. In this work, we propose a universal cross-modality knowledge distillation framework (UniDistill) to improve the performance of single-modality detectors. Specifically, during training, UniDistill projects the features of both the teacher and the student detector into Bird's-Eye-View (BEV), which is a friendly representation for different modalities. Then, three distillation losses are calculated to sparsely align the foreground features, helping the student learn from the teacher without introducing additional cost during inference. Taking advantage of the similar detection paradigm of different detectors in BEV, UniDistill easily supports LiDAR-to-camera, camera-to-LiDAR, fusion-to-LiDAR and fusion-to-camera distillation paths. Furthermore, the three distillation losses can filter the effect of misaligned background information and balance between objects of different sizes, improving the distillation effectiveness. Extensive experiments on nuScenes demonstrate that UniDistill effectively improves the mAP and NDS of student detectors by 2.0% 3.2%.
****
SemiCVT: Semi-Supervised Convolutional Vision Transformer for Semantic Segmentation
Huimin Huang, Shiao Xie, Lanfen Lin, Ruofeng Tong, Yen-Wei Chen, Yuexiang Li, Hong Wang, Yawen Huang, Yefeng Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11340-11349
Semi-supervised learning improves data efficiency of deep models by leveraging unlabeled samples to alleviate the reliance on a large set of labeled samples. These successes concentrate on the pixel-wise consistency by using convolutional neural networks (CNNs) but fail to address both global learning capability and class-level features for unlabeled data. Recent works raise a new trend that Trans - former achieves superior performance on the entire feature map in various tasks. In this paper, we unify the current dominant Mean-Teacher approaches by reconciling intra- model and inter-model properties for semi-supervised segmentation to produce a novel algorithm, SemiCVT, that absorbs the quintessence of CNNs and Transformer in a comprehensive way. Specifically, we first design a parallel CNN-Transformer architecture (CVT) with introducing an intra-model local-global interaction schema (LGI) in Fourier domain for full integration. The inter-model class- wise consistency is further presented to complement the class-level statistics of CNNs and Transformer in a cross- teaching manner. Extensive empirical evidence shows that SemiCVT yields consistent improvements over the state-of- the-art methods in two public benchmarks.
****
Fine-Tuned CLIP Models Are Efficient Video Learners
Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6545-6554
Large-scale multi-modal training with image-text pairs imparts strong generalization to CLIP model. Since training on a similar scale for videos is infeasible, recent approaches focus on the effective transfer of image-based CLIP to the video domain. In this pursuit, new parametric modules are added to learn temporal information and inter-frame relationships which require meticulous design efforts. Furthermore, when the resulting models are learned on videos, they tend to overfit on the given task distribution and lack in generalization aspect. This begs the following question: How to effectively transfer image-level CLIP representations to videos? In this work, we show that a simple Video Fine-tuned CLIP (ViFi-CLIP) baseline is generally sufficient to bridge the domain gap from images to videos. Our qualitative analysis illustrates that the frame-level processing from CLIP image-encoder followed by feature pooling and similarity matching with corresponding text embeddings helps in implicitly modeling the temporal cues within ViFi-CLIP. Such fine-tuning helps the model to focus on scene dynamics, moving objects and inter-object relationships. For low-data regimes where full fine-tuning is not viable, we propose a 'bridge and prompt' approach that first uses fi

netuning to bridge the domain gap and then learns prompts on language and vision side to adapt CLIP representations. We extensively evaluate this simple yet strong baseline on zero-shot, base-to-novel generalization, few-shot and fully supervised settings across five video benchmarks. Our code and models will be publicly released.
****
Towards Open-World Segmentation of Parts
Tai-Yu Pan, Qing Liu, Wei-Lun Chao, Brian Price; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15392-15401
Segmenting object parts such as cup handles and animal bodies is important in many real-world applications but requires more annotation effort. The largest dataset nowadays contains merely two hundred object categories, implying the difficulty to scale up part segmentation to an unconstrained setting. To address this, we propose to explore a seemingly simplified but empirically useful and scalable task, class-agnostic part segmentation. In this problem, we disregard the part class labels in training and instead treat all of them as a single part class. We argue and demonstrate that models trained without part classes can better localize parts and segment them on objects unseen in training. We then present two further improvements. First, we propose to make the model object-aware, leveraging the fact that parts are "compositions" whose extents are bounded by objects, whose appearances are by nature not independent but bundled. Second, we introduce a novel approach to improve part segmentation on unseen objects, inspired by an interesting finding --- for unseen objects, the pixel-wise features extracted by the model often reveal high-quality part segments. To this end, we propose a novel self-supervised procedure that iterates between pixel clustering and supervised contrastive learning that pulls pixels closer or pushes them away. Via extensive experiments on PartImageNet and Pascal-Part, we show notable and consistent gains by our approach, essentially a critical step towards open-world part segmentation.
****
Stitchable Neural Networks
Zizheng Pan, Jianfei Cai, Bohan Zhuang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16102-16112
The public model zoo containing enormous powerful pretrained model families (e.g., ResNet/DeiT) has reached an unprecedented scope than ever, which significantly contributes to the success of deep learning. As each model family consists of pretrained models with diverse scales (e.g., DeiT-Ti/S/B), it naturally arises a fundamental question of how to efficiently assemble these readily available models in a family for dynamic accuracy-efficiency trade-offs at runtime. To this end, we present Stitchable Neural Networks (SN-Net), a novel scalable and efficient framework for model deployment. It cheaply produces numerous networks with different complexity and performance trade-offs given a family of pretrained neural networks, which we call anchors. Specifically, SN-Net splits the anchors across the blocks/layers and then stitches them together with simple stitching layers to map the activations from one anchor to another. With only a few epochs of training, SN-Net effectively interpolates between the performance of anchors with varying scales. At runtime, SN-Net can instantly adapt to dynamic resource constraints by switching the stitching positions. Extensive experiments on ImageNet classification demonstrate that SN-Net can obtain on-par or even better performance than many individually trained networks while supporting diverse deployment scenarios. For example, by stitching Swin Transformers, we challenge hundreds of models in Timm model zoo with a single network. We believe this new elastic model framework can serve as a strong baseline for further research in wider communities.
****
Collaborative Diffusion for Multi-Modal Face Generation and Editing
Ziqi Huang, Kelvin C.K. Chan, Yuming Jiang, Ziwei Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6080-6090
Diffusion models arise as a powerful generative tool recently. Despite the great

progress, existing diffusion models mainly focus on uni-modal control, i.e., the diffusion process is driven by only one modality of condition. To further unleash the users' creativity, it is desirable for the model to be controllable by multiple modalities simultaneously, e.g., generating and editing faces by describing the age (text-driven) while drawing the face shape (mask-driven). In this work, we present Collaborative Diffusion, where pre-trained uni-modal diffusion models collaborate to achieve multi-modal face generation and editing without re-training. Our key insight is that diffusion models driven by different modalities are inherently complementary regarding the latent denoising steps, where bilateral connections can be established upon. Specifically, we propose dynamic diffuser, a meta-network that adaptively hallucinates multi-modal denoising steps by predicting the spatial-temporal influence functions for each pre-trained uni-modal model. Collaborative Diffusion not only collaborates generation capabilities from uni-modal diffusion models, but also integrates multiple uni-modal manipulations to perform multi-modal editing. Extensive qualitative and quantitative experiments demonstrate the superiority of our framework in both image quality and condition consistency.
****
DejaVu: Conditional Regenerative Learning To Enhance Dense Prediction
Shubhankar Borse, Debasmit Das, Hyojin Park, Hong Cai, Risheek Garrepalli, Fatih Porikli; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19466-19477
We present DejaVu, a novel framework which leverages conditional image regeneration as additional supervision during training to improve deep networks for dense prediction tasks such as segmentation, depth estimation, and surface normal prediction. First, we apply redaction to the input image, which removes certain structural information by sparse sampling or selective frequency removal. Next, we use a conditional regenerator, which takes the redacted image and the dense predictions as inputs, and reconstructs the original image by filling in the missing structural information. In the redacted image, structural attributes like boundaries are broken while semantic context is largely preserved. In order to make the regeneration feasible, the conditional generator will then require the structure information from the other input source, i.e., the dense predictions. As such, by including this conditional regeneration objective during training, DejaVu encourages the base network to learn to embed accurate scene structure in its dense prediction. This leads to more accurate predictions with clearer boundaries and better spatial consistency. When it is feasible to leverage additional computation, DejaVu can be extended to incorporate an attention-based regeneration module within the dense prediction network, which further improves accuracy. Through extensive experiments on multiple dense prediction benchmarks such as Cityscapes, COCO, ADE20K, NYUD-v2, and KITTI, we demonstrate the efficacy of employing DejaVu during training, as it outperforms SOTA methods at no added computation cost.
****
MACARONS: Mapping and Coverage Anticipation With RGB Online Self-Supervision
Antoine Guédon, Tom Monnier, Pascal Monasse, Vincent Lepetit; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 940-951
We introduce a method that simultaneously learns to explore new large environments and to reconstruct them in 3D from color images only. This is closely related to the Next Best View problem (NBV), where one has to identify where to move the camera next to improve the coverage of an unknown scene. However, most of the current NBV methods rely on depth sensors, need 3D supervision and/or do not scale to large scenes. Our method requires only a color camera and no 3D supervision. It simultaneously learns in a self-supervised fashion to predict a volume occupancy field from color images and, from this field, to predict the NBV. Thanks to this approach, our method performs well on new scenes as it is not biased towards any training 3D data. We demonstrate this on a recent dataset made of various 3D scenes and show it performs even better than recent methods requiring a depth sensor, which is not a realistic assumption for outdoor scenes captured with

a flying drone.
****

Audio-Visual Grouping Network for Sound Localization From Mixtures
Shentong Mo, Yapeng Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10565-10574
Sound source localization is a typical and challenging task that predicts the location of sound sources in a video. Previous single-source methods mainly used the audio-visual association as clues to localize sounding objects in each frame. Due to the mixed property of multiple sound sources in the original space, there exist rare multi-source approaches to localizing multiple sources simultaneously, except for one recent work using a contrastive random walk in the graph with images and separated sound as nodes. Despite their promising performance, they can only handle a fixed number of sources, and they cannot learn compact class-aware representations for individual sources. To alleviate this shortcoming, in this paper, we propose a novel audio-visual grouping network, namely AVGN, that can directly learn category-wise semantic features for each source from the input audio mixture and frame to localize multiple sources simultaneously. Specifically, our AVGN leverages learnable audio-visual class tokens to aggregate class-aware source features. Then, the aggregated semantic features for each source can be used as guidance to localize the corresponding visual regions. Compared to existing multi-source methods, our new framework can localize a flexible number of sources and disentangle category-aware audio-visual representations for individual sound sources. We conduct extensive experiments on MUSIC, VGGSound-Instruments, and VGG-Sound Sources benchmarks. The results demonstrate that the proposed AVGN can achieve state-of-the-art sounding object localization performance on both single-source and multi-source scenarios.
****

Fair Federated Medical Image Segmentation via Client Contribution Estimation
Meirui Jiang, Holger R. Roth, Wenqi Li, Dong Yang, Can Zhao, Vishwesh Nath, Daguang Xu, Qi Dou, Ziyue Xu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16302-16311
How to ensure fairness is an important topic in federated learning (FL). Recent studies have investigated how to reward clients based on their contribution (collaboration fairness), and how to achieve uniformity of performance across clients (performance fairness). Despite achieving progress on either one, we argue that it is critical to consider them together, in order to engage and motivate more diverse clients joining FL to derive a high-quality global model. In this work, we propose a novel method to optimize both types of fairness simultaneously. Specifically, we propose to estimate client contribution in gradient and data space. In gradient space, we monitor the gradient direction differences of each client with respect to others. And in data space, we measure the prediction error on client data using an auxiliary model. Based on this contribution estimation, we propose a FL method, federated training via contribution estimation (FedCE), i.e., using estimation as global model aggregation weights. We have theoretically analyzed our method and empirically evaluated it on two real-world medical datasets. The effectiveness of our approach has been validated with significant performance improvements, better collaboration fairness, better performance fairness, and comprehensive analytical studies.
****

Dynamic Generative Targeted Attacks With Pattern Injection
Weiwei Feng, Nanqing Xu, Tianzhu Zhang, Yongdong Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16404-16414
Adversarial attacks can evaluate model robustness and have been of great concerns in recent years. Among various attacks, targeted attacks aim at misleading victim models to output adversary-desired predictions, which are more challenging and threatening than untargeted ones. Existing targeted attacks can be roughly divided into instancespecific and instance-agnostic attacks. Instance-specific attacks craft adversarial examples via iterative gradient updating on the specific instance. In contrast, instanceagnostic attacks learn a universal perturbation o

r a generative model on the global dataset to perform attacks. However they rely too much on the classification boundary of substitute models, ignoring the realistic distribution of target class, which may result in limited targeted attack performance. And there is no attempt to simultaneously combine the information of the specific instance and the global dataset. To deal with these limitations, we first conduct an analysis via a causal graph and propose to craft transferable targeted adversarial examples by injecting target patterns. Based on this analysis, we introduce a generative attack model composed of a cross-attention guided convolution module and a pattern injection module. Concretely, the former adopts a dynamic convolution kernel and a static convolution kernel for the specific instance and the global dataset, respectively, which can inherit the advantages of both instance-specific and instance-agnostic attacks. And the pattern injection module utilizes a pattern prototype to encode target patterns, which can guide the generation of targeted adversarial examples. Besides, we also provide rigorous theoretical analysis to guarantee the effectiveness of our method. Extensive experiments demonstrate that our method show superior performance than 10 existing adversarial attacks against 13 models.
****

Tracking Multiple Deformable Objects in Egocentric Videos

Mingzhen Huang, Xiaoxing Li, Jun Hu, Honghong Peng, Siwei Lyu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1461-1471

Most existing multiple object tracking (MOT) methods that solely rely on appearance features struggle in tracking highly deformable objects. Other MOT methods that use motion clues to associate identities across frames have difficulty handling egocentric videos effectively or efficiently. In this work, we propose DETracker, a new MOT method that jointly detects and tracks deformable objects in egocentric videos. DETracker uses three novel modules, namely the motion disentanglement network (MDN), the patch association network (PAN) and the patch memory network (PMN), to explicitly tackle the difficulties caused by severe ego motion and fast morphing target objects. DETracker is end-to-end trainable and achieves near real-time speed. We also present DogThruGlasses, a large-scale deformable multi-object tracking dataset, with 150 videos and 73K annotated frames, collected by smart glasses. DETracker outperforms existing state-of-the-art method on the DogThruGlasses dataset and YouTube-Hand dataset.
****

Visual Recognition by Request

Chufeng Tang, Lingxi Xie, Xiaopeng Zhang, Xiaolin Hu, Qi Tian; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15265-15274

Humans have the ability of recognizing visual semantics in an unlimited granularity, but existing visual recognition algorithms cannot achieve this goal. In this paper, we establish a new paradigm named visual recognition by request (ViRReq) to bridge the gap. The key lies in decomposing visual recognition into atomic tasks named requests and leveraging a knowledge base, a hierarchical and text-based dictionary, to assist task definition. ViRReq allows for (i) learning complicated whole-part hierarchies from highly incomplete annotations and (ii) inserting new concepts with minimal efforts. We also establish a solid baseline by integrating language-driven recognition into recent semantic and instance segmentation methods, and demonstrate its flexible recognition ability on CPP and ADE20K, two datasets with hierarchical whole-part annotations.
****

SmartBrush: Text and Shape Guided Object Inpainting With Diffusion Model

Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, Kun Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22428-22437

Generic image inpainting aims to complete a corrupted image by borrowing surrounding information, which barely generates novel content. By contrast, multi-modal inpainting provides more flexible and useful controls on the inpainted content, e.g., a text prompt can be used to describe an object with richer attributes, a

nd a mask can be used to constrain the shape of the inpainted object rather than being only considered as a missing area. We propose a new diffusion-based model named SmartBrush for completing a missing region with an object using both text and shape-guidance. While previous work such as DALLE-2 and Stable Diffusion can do text-guided inapinting they do not support shape guidance and tend to modify background texture surrounding the generated object. Our model incorporates both text and shape guidance with precision control. To preserve the background better, we propose a novel training and sampling strategy by augmenting the diffusion U-net with object-mask prediction. Lastly, we introduce a multi-task training strategy by jointly training inpainting with text-to-image generation to leverage more training data. We conduct extensive experiments showing that our model outperforms all baselines in terms of visual quality, mask controllability, and background preservation.
****
REC-MV: REconstructing 3D Dynamic Cloth From Monocular Videos
Lingteng Qiu, Guanying Chen, Jiapeng Zhou, Mutian Xu, Junle Wang, Xiaoguang Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4637-4646
Reconstructing dynamic 3D garment surfaces with open boundaries from monocular videos is an important problem as it provides a practical and low-cost solution for clothes digitization. Recent neural rendering methods achieve high-quality dynamic clothed human reconstruction results from monocular video, but these methods cannot separate the garment surface from the body. Moreover, despite existing garment reconstruction methods based on feature curve representation demonstrating impressive results for garment reconstruction from a single image, they struggle to generate temporally consistent surfaces for the video input. To address the above limitations, in this paper, we formulate this task as an optimization problem of 3D garment feature curves and surface reconstruction from monocular video. We introduce a novel approach, called REC-MV to jointly optimize the explicit feature curves and the implicit signed distance field (SDF) of the garments. Then the open garment meshes can be extracted via garment template registration in the canonical space. Experiments on multiple casually captured datasets show that our approach outperforms existing methods and can produce high-quality dynamic garment surfaces.
****
JRDB-Pose: A Large-Scale Dataset for Multi-Person Pose Estimation and Tracking
Edward Vendrow, Duy Tho Le, Jianfei Cai, Hamid Rezatofighi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4811-4820
Autonomous robotic systems operating in human environments must understand their surroundings to make accurate and safe decisions. In crowded human scenes with close-up human-robot interaction and robot navigation, a deep understanding of surrounding people requires reasoning about human motion and body dynamics over time with human body pose estimation and tracking. However, existing datasets captured from robot platforms either do not provide pose annotations or do not reflect the scene distribution of social robots. In this paper, we introduce JRDB-Pose, a large-scale dataset and benchmark for multi-person pose estimation and tracking. JRDB-Pose extends the existing JRDB which includes videos captured from a social navigation robot in a university campus environment, containing challenging scenes with crowded indoor and outdoor locations and a diverse range of scales and occlusion types. JRDB-Pose provides human pose annotations with per-keypoint occlusion labels and track IDs consistent across the scene and with existing annotations in JRDB. We conduct a thorough experimental study of state-of-the-art multi-person pose estimation and tracking methods on JRDB-Pose, showing that our dataset imposes new challenges for the existing methods. JRDB-Pose is available at https://jrdb.erc.monash.edu/.
****
AsyFOD: An Asymmetric Adaptation Paradigm for Few-Shot Domain Adaptive Object Detection
Yipeng Gao, Kun-Yu Lin, Junkai Yan, Yaowei Wang, Wei-Shi Zheng; Proceedings of t

he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3261-3271
In this work, we study few-shot domain adaptive object detection (FSDAOD), where only a few target labeled images are available for training in addition to sufficient source labeled images. Critically, in FSDAOD, the data-scarcity in the target domain leads to an extreme data imbalance between the source and target domains, which potentially causes over-adaptation in traditional feature alignment. To address the data imbalance problem, we propose an asymmetric adaptation paradigm, namely AsyFOD, which leverages the source and target instances from different perspectives. Specifically, by using target distribution estimation, the AsyFOD first identifies the target-similar source instances, which serves for augmenting the limited target instances. Then, we conduct asynchronous alignment between target-dissimilar source instances and augmented target instances, which is simple yet effective for alleviating the over-adaptation. Extensive experiments demonstrate that the proposed AsyFOD outperforms all state-of-the-art methods on four FSDAOD benchmarks with various environmental variances, e.g., 3.1% mAP improvement on Cityscapes-to-FoggyCityscapes and 2.9% mAP increase on Sim10k-to-Cityscapes. The code is available at https://github.com/Hlings/AsyFOD.
****
RUST: Latent Neural Scene Representations From Unposed Imagery
Mehdi S. M. Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lu■i■, Klaus Greff; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17297-17306
Inferring the structure of 3D scenes from 2D observations is a fundamental challenge in computer vision. Recently popularized approaches based on neural scene representations have achieved tremendous impact and have been applied across a variety of applications. One of the major remaining challenges in this space is training a single model which can provide latent representations which effectively generalize beyond a single scene. Scene Representation Transformer (SRT) has shown promise in this direction, but scaling it to a larger set of diverse scenes is challenging and necessitates accurately posed ground truth data. To address this problem, we propose RUST (Really Unposed Scene representation Transformer), a pose-free approach to novel view synthesis trained on RGB images alone. Our main insight is that one can train a Pose Encoder that peeks at the target image and learns a latent pose embedding which is used by the decoder for view synthesis. We perform an empirical investigation into the learned latent pose structure and show that it allows meaningful test-time camera transformations and accurate explicit pose readouts. Perhaps surprisingly, RUST achieves similar quality as methods which have access to perfect camera pose, thereby unlocking the potential for large-scale training of amortized neural scene representations.
****
PointCert: Point Cloud Classification With Deterministic Certified Robustness Guarantees
Jinghuai Zhang, Jinyuan Jia, Hongbin Liu, Neil Zhenqiang Gong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9496-9505
Point cloud classification is an essential component in many security-critical applications such as autonomous driving and augmented reality. However, point cloud classifiers are vulnerable to adversarially perturbed point clouds. Existing certified defenses against adversarial point clouds suffer from a key limitation: their certified robustness guarantees are probabilistic, i.e., they produce an incorrect certified robustness guarantee with some probability. In this work, we propose a general framework, namely PointCert, that can transform an arbitrary point cloud classifier to be certifiably robust against adversarial point clouds with deterministic guarantees. PointCert certifiably predicts the same label for a point cloud when the number of arbitrarily added, deleted, and/or modified points is less than a threshold. Moreover, we propose multiple methods to optimize the certified robustness guarantees of PointCert in three application scenarios. We systematically evaluate PointCert on ModelNet and ScanObjectNN benchmark datasets. Our results show that PointCert substantially outperforms state-of-the

-art certified defenses even though their robustness guarantees are probabilistic.
****

Open Set Action Recognition via Multi-Label Evidential Learning
Chen Zhao, Dawei Du, Anthony Hoogs, Christopher Funk; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22982-22991
Existing methods for open set action recognition focus on novelty detection that assumes video clips show a single action, which is unrealistic in the real world. We propose a new method for open set action recognition and novelty detection via MUlti-Label Evidential learning (MULE), that goes beyond previous novel action detection methods by addressing the more general problems of single or multiple actors in the same scene, with simultaneous action(s) by any actor. Our Beta Evidential Neural Network estimates multi-action uncertainty with Beta densities based on actor-context-object relation representations. An evidence debiasing constraint is added to the objective func- tion for optimization to reduce the static bias of video representations, which can incorrectly correlate predictions and static cues. We develop a primal-dual average scheme update-based learning algorithm to optimize the proposed problem and provide corresponding theoretical analysis. Besides, uncertainty and belief-based novelty estimation mechanisms are formulated to detect novel actions. Extensive experiments on two real-world video datasets show that our proposed approach achieves promising performance in single/multi-actor, single/multi-action settings. Our code and models are released at https://github.com/charliezhaoyinpeng/mule.
****

MAP: Multimodal Uncertainty-Aware Vision-Language Pre-Training Model
Yatai Ji, Junjie Wang, Yuan Gong, Lin Zhang, Yanru Zhu, Hongfa Wang, Jiaxing Zhang, Tetsuya Sakai, Yujiu Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23262-23271
Multimodal semantic understanding often has to deal with uncertainty, which means the obtained messages tend to refer to multiple targets. Such uncertainty is problematic for our interpretation, including inter- and intra-modal uncertainty. Little effort has studied the modeling of this uncertainty, particularly in pre-training on unlabeled datasets and fine-tuning in task-specific downstream datasets. In this paper, we project the representations of all modalities as probabilistic distributions via a Probability Distribution Encoder (PDE) by utilizing sequence-level interactions. Compared to the exiting deterministic methods, such uncertainty modeling can convey richer multimodal semantic information and more complex relationships. Furthermore, we integrate uncertainty modeling with popular pre-training frameworks and propose suitable pre-training tasks: Distribution-based Vision-Language Contrastive learning (D-VLC), Distribution-based Masked Language Modeling (D-MLM), and Distribution-based Image-Text Matching (D-ITM). The fine-tuned models are applied to challenging downstream tasks, including image-text retrieval, visual question answering, visual reasoning, and visual entailment, and achieve state-of-the-art results.
****

DualRel: Semi-Supervised Mitochondria Segmentation From a Prototype Perspective
Huayu Mai, Rui Sun, Tianzhu Zhang, Zhiwei Xiong, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19617-19626
Automatic mitochondria segmentation enjoys great popularity with the development of deep learning. However, existing methods rely heavily on the labor-intensive manual gathering by experienced domain experts. And naively applying semi-supervised segmentation methods in the natural image field to mitigate the labeling cost is undesirable. In this work, we analyze the gap between mitochondrial images and natural images and rethink how to achieve effective semi-supervised mitochondria segmentation, from the perspective of reliable prototype-level supervision. We propose a novel end-to-end dual-reliable (DualRel) network, including a reliable pixel aggregation module and a reliable prototype selection module. The proposed DualRel enjoys several merits. First, to learn the prototypes well witho

ut any explicit supervision, we carefully design the referential correlation to rectify the direct pairwise correlation. Second, the reliable prototype selection module is responsible for further evaluating the reliability of prototypes in constructing prototype-level consistency regularization. Extensive experimental results on three challenging benchmarks demonstrate that our method performs favorably against state-of-the-art semi-supervised segmentation methods. Importantly, with extremely few samples used for training, DualRel is also on par with current state-of-the-art fully supervised methods.
****
Federated Learning With Data-Agnostic Distribution Fusion
Jian-hui Duan, Wenzhong Li, Derun Zou, Ruichen Li, Sanglu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8074-8083
Federated learning has emerged as a promising distributed machine learning paradigm to preserve data privacy. One of the fundamental challenges of federated learning is that data samples across clients are usually not independent and identically distributed (non-IID), leading to slow convergence and severe performance drop of the aggregated global model. To facilitate model aggregation on non-IID data, it is desirable to infer the unknown global distributions without violating privacy protection policy. In this paper, we propose a novel data-agnostic distribution fusion based model aggregation method called FedFusion to optimize federated learning with non-IID local datasets, based on which the heterogeneous clients' data distributions can be represented by a global distribution of several virtual fusion components with different parameters and weights. We develop a Variational AutoEncoder (VAE) method to learn the optimal parameters of the distribution fusion components based on limited statistical information extracted from the local models, and apply the derived distribution fusion model to optimize federated model aggregation with non-IID data. Extensive experiments based on various federated learning scenarios with real-world datasets show that FedFusion achieves significant performance improvement compared to the state-of-the-art.
****
Cap4Video: What Can Auxiliary Captions Do for Text-Video Retrieval?
Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10704-10713
Most existing text-video retrieval methods focus on cross-modal matching between the visual content of videos and textual query sentences. However, in real-world scenarios, online videos are often accompanied by relevant text information such as titles, tags, and even subtitles, which can be utilized to match textual queries. This insight has motivated us to propose a novel approach to text-video retrieval, where we directly generate associated captions from videos using zero-shot video captioning with knowledge from web-scale pre-trained models (e.g., CLIP and GPT-2). Given the generated captions, a natural question arises: what benefits do they bring to text-video retrieval? To answer this, we introduce Cap4Video, a new framework that leverages captions in three ways: i) Input data: video-caption pairs can augment the training data. ii) Intermediate feature interaction: we perform cross-modal feature interaction between the video and caption to produce enhanced video representations. iii) Output score: the Query-Caption matching branch can complement the original Query-Video matching branch for text-video retrieval. We conduct comprehensive ablation studies to demonstrate the effectiveness of our approach. Without any post-processing, Cap4Video achieves state-of-the-art performance on four standard text-video retrieval benchmarks: MSR-VTT (51.4%), VATEX (66.6%), MSVD (51.8%), and DiDeMo (52.0%). The code is available at https://github.com/whwu95/Cap4Video.
****
Progressive Semantic-Visual Mutual Adaption for Generalized Zero-Shot Learning
Man Liu, Feng Li, Chunjie Zhang, Yunchao Wei, Huihui Bai, Yao Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15337-15346
Generalized Zero-Shot Learning (GZSL) identifies unseen categories by knowledge

transferred from the seen domain, relying on the intrinsic interactions between visual and semantic information. Prior works mainly localize regions correspondi ng to the sharing attributes. When various visual appearances correspond to the same attribute, the sharing attributes inevitably introduce semantic ambiguity, hampering the exploration of accurate semantic-visual interactions. In this pape r, we deploy the dual semantic-visual transformer module (DSVTM) to progressivel y model the correspondences between attribute prototypes and visual features, co nstituting a progressive semantic-visual mutual adaption (PSVMA) network for sem antic disambiguation and knowledge transferability improvement. Specifically, DS VTM devises an instance-motivated semantic encoder that learns instance-centric prototypes to adapt to different images, enabling the recast of the unmatched se mantic-visual pair into the matched one. Then, a semantic-motivated instance dec oder strengthens accurate cross-domain interactions between the matched pair for semantic-related instance adaption, encouraging the generation of unambiguous v isual representations. Moreover, to mitigate the bias towards seen classes in GZ SL, a debiasing loss is proposed to pursue response consistency between seen and unseen predictions. The PSVMA consistently yields superior performances against other state-of-the-art methods. Code will be available at: https://github.com/M anLiuCoder/PSVMA.
****
Gated Multi-Resolution Transfer Network for Burst Restoration and Enhancement
Nancy Mehta, Akshay Dudhane, Subrahmanyam Murala, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF Conference on Computer Vision a nd Pattern Recognition (CVPR), 2023, pp. 22201-22210
Burst image processing is becoming increasingly popular in recent years. However , it is a challenging task since individual burst images undergo multiple degrad ations and often have mutual misalignments resulting in ghosting and zipper arti facts. Existing burst restoration methods usually do not consider the mutual cor relation and non-local contextual information among burst frames, which tends to limit these approaches in challenging cases. Another key challenge lies in the robust up-sampling of burst frames. The existing up-sampling methods cannot effe ctively utilize the advantages of single-stage and progressive up-sampling strat egies with conventional and/or recent up-samplers at the same time. To address t hese challenges, we propose a novel Gated Multi-Resolution Transfer Network (GMT Net) to reconstruct a spatially precise high-quality image from a burst of low-q uality raw images. GMTNet consists of three modules optimized for burst processi ng tasks: Multi-scale Burst Feature Alignment (MBFA) for feature denoising and a lignment, Transposed-Attention Feature Merging (TAFM) for multi-frame feature ag gregation, and Resolution Transfer Feature Up-sampler (RTFU) to up-scale merged features and construct a high-quality output image. Detailed experimental analys is on five datasets validate our approach and sets a new state-of-the-art for bu rst super-resolution, burst denoising, and low-light burst enhancement. Our code s and models are available at https://github.com/nanmehta/GMTNet.
****
Improving Commonsense in Vision-Language Models via Knowledge Graph Riddles
Shuquan Ye, Yujia Xie, Dongdong Chen, Yichong Xu, Lu Yuan, Chenguang Zhu, Jing L iao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recog nition (CVPR), 2023, pp. 2634-2645
This paper focuses on analyzing and improving the commonsense ability of recent popular vision-language (VL) models. Despite the great success, we observe that existing VL-models still lack commonsense knowledge/reasoning ability (e.g., "Le mons are sour"), which is a vital component towards artificial general intellige nce. Through our analysis, we find one important reason is that existing large-s cale VL datasets do not contain much commonsense knowledge, which motivates us t o improve the commonsense of VL-models from the data perspective. Rather than co llecting a new VL training dataset, we propose a more scalable strategy, i.e., " Data Augmentation with kNowledge graph linearization for CommonsensE capability" (DANCE). It can be viewed as one type of data augmentation technique, which can inject commonsense knowledge into existing VL datasets on the fly during traini ng. More specifically, we leverage the commonsense knowledge graph (e.g., Concep

tNet) and create variants of text description in VL datasets via bidirectional sub-graph sequentialization. For better commonsense evaluation, we further propose the first retrieval-based commonsense diagnostic benchmark. By conducting extensive experiments on some representative VL-models, we demonstrate that our DANCE technique is able to significantly improve the commonsense ability while maintaining the performance on vanilla retrieval tasks.
****
S3C: Semi-Supervised VQA Natural Language Explanation via Self-Critical Learning
Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao, Peng Wang, Yanning Zhang, Qi Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2646-2656
VQA Natural Language Explanation (VQA-NLE) task aims to explain the decision-making process of VQA models in natural language. Unlike traditional attention or gradient analysis, free-text rationales can be easier to understand and gain users' trust. Existing methods mostly use post-hoc or self-rationalization models to obtain a plausible explanation. However, these frameworks are bottlenecked by the following challenges: 1) the reasoning process cannot be faithfully responded to and suffer from the problem of logical inconsistency. 2) Human-annotated explanations are expensive and time-consuming to collect. In this paper, we propose a new Semi-Supervised VQA-NLE via Self-Critical Learning (S3C), which evaluates the candidate explanations by answering rewards to improve the logical consistency between answers and rationales. With a semi-supervised learning framework, the S3C can benefit from a tremendous amount of samples without human-annotated explanations. A large number of automatic measures and human evaluations all show the effectiveness of our method. Meanwhile, the framework achieves a new state-of-the-art performance on the two VQA-NLE datasets.
****
Spatio-Focal Bidirectional Disparity Estimation From a Dual-Pixel Image
Donggun Kim, Hyeonjoong Jang, Inchul Kim, Min H. Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5023-5032
Dual-pixel photography is monocular RGB-D photography with an ultra-high resolution, enabling many applications in computational photography. However, there are still several challenges to fully utilizing dual-pixel photography. Unlike the conventional stereo pair, the dual pixel exhibits a bidirectional disparity that includes positive and negative values, depending on the focus plane depth in an image. Furthermore, capturing a wide range of dual-pixel disparity requires a shallow depth of field, resulting in a severely blurred image, degrading depth estimation performance. Recently, several data-driven approaches have been proposed to mitigate these two challenges. However, due to the lack of the ground-truth dataset of the dual-pixel disparity, existing data-driven methods estimate either inverse depth or blurriness map. In this work, we propose a self-supervised learning method that learns bidirectional disparity by utilizing the nature of anisotropic blur kernels in dual-pixel photography. We observe that the dual-pixel left/right images have reflective-symmetric anisotropic kernels, so their sum is equivalent to that of a conventional image. We take a self-supervised training approach with the novel kernel-split symmetry loss accounting for the phenomenon. Our method does not rely on a training dataset of dual-pixel disparity that does not exist yet. Our method can estimate a complete disparity map with respect to the focus-plane depth from a dual-pixel image, outperforming the baseline dual-pixel methods.
****
Block Selection Method for Using Feature Norm in Out-of-Distribution Detection
Yeonguk Yu, Sungho Shin, Seongju Lee, Changhyun Jun, Kyoobin Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15701-15711
Detecting out-of-distribution (OOD) inputs during the inference stage is crucial for deploying neural networks in the real world. Previous methods commonly relied on the output of a network derived from the highly activated feature map. In this study, we first revealed that a norm of the feature map obtained from the o

ther block than the last block can be a better indicator of OOD detection. Motivated by this, we propose a simple framework consisting of FeatureNorm: a norm of the feature map and NormRatio: a ratio of FeatureNorm for ID and OOD to measure the OOD detection performance of each block. In particular, to select the block that provides the largest difference between FeatureNorm of ID and FeatureNorm of OOD, we create jigsaw puzzles as pseudo OOD from ID training samples and calculate NormRatio, and the block with the largest value is selected. After the suitable block is selected, OOD detection with the FeatureNorm outperforms other OOD detection methods by reducing FPR95 by up to 52.77% on CIFAR10 benchmark and by up to 48.53% on ImageNet benchmark. We demonstrate that our framework can generalize to various architectures and the importance of block selection, which can improve previous OOD detection methods as well.
****
PIDNet: A Real-Time Semantic Segmentation Network Inspired by PID Controllers
Jiacong Xu, Zixiang Xiong, Shankar P. Bhattacharyya; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19529-19539
Two-branch network architecture has shown its efficiency and effectiveness in real-time semantic segmentation tasks. However, direct fusion of high-resolution details and low-frequency context has the drawback of detailed features being easily overwhelmed by surrounding contextual information. This overshoot phenomenon limits the improvement of the segmentation accuracy of existing two-branch models. In this paper, we make a connection between Convolutional Neural Networks (CNN) and Proportional-Integral-Derivative (PID) controllers and reveal that a two-branch network is equivalent to a Proportional-Integral (PI) controller, which inherently suffers from similar overshoot issues. To alleviate this problem, we propose a novel three-branch network architecture: PIDNet, which contains three branches to parse detailed, context and boundary information, respectively, and employs boundary attention to guide the fusion of detailed and context branches. Our family of PIDNets achieve the best trade-off between inference speed and accuracy and their accuracy surpasses all the existing models with similar inference speed on the Cityscapes and CamVid datasets. Specifically, PIDNet-S achieves 78.6 mIOU with inference speed of 93.2 FPS on Cityscapes and 80.1 mIOU with speed of 153.7 FPS on CamVid.
****
Four-View Geometry With Unknown Radial Distortion
Petr Hruby, Viktor Korotynskiy, Timothy Duff, Luke Oeding, Marc Pollefeys, Tomas Pajdla, Viktor Larsson; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8990-9000
We present novel solutions to previously unsolved problems of relative pose estimation from images whose calibration parameters, namely focal lengths and radial distortion, are unknown. Our approach enables metric reconstruction without modeling these parameters. The minimal case for reconstruction requires 13 points in 4 views for both the calibrated and uncalibrated cameras. We describe and implement the first solution to these minimal problems. In the calibrated case, this may be modeled as a polynomial system of equations with 3584 solutions. Despite the apparent intractability, the problem decomposes spectacularly. Each solution falls into a Euclidean symmetry class of size 16, and we can estimate 224 class representatives by solving a sequence of three subproblems with 28, 2, and 4 solutions. We highlight the relationship between internal constraints on the radial quadrifocal tensor and the relations among the principal minors of a 4x4 matrix. We also address the case of 4 upright cameras, where 7 points are minimal. Finally, we evaluate our approach on simulated and real data and benchmark against previous calibration-free solutions, and show that our method provides an efficient startup for an SfM pipeline with radial cameras.
****
Rethinking Optical Flow From Geometric Matching Consistent Perspective
Qiaole Dong, Chenjie Cao, Yanwei Fu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1337-1347
Optical flow estimation is a challenging problem remaining unsolved. Recent deep

learning based optical flow models have achieved considerable success. However, these models often train networks from the scratch on standard optical flow data, which restricts their ability to robustly and geometrically match image features. In this paper, we propose a rethinking to previous optical flow estimation. We particularly leverage Geometric Image Matching (GIM) as a pre-training task for the optical flow estimation (MatchFlow) with better feature representations, as GIM shares some common challenges as optical flow estimation, and with massive labeled real-world data. Thus, matching static scenes helps to learn more fundamental feature correlations of objects and scenes with consistent displacements. Specifically, the proposed MatchFlow model employs a QuadTree attention-based network pre-trained on MegaDepth to extract coarse features for further flow regression. Extensive experiments show that our model has great cross-dataset generalization. Our method achieves 11.5% and 10.1% error reduction from GMA on Sintel clean pass and KITTI test set. At the time of anonymous submission, our MatchFlow(G) enjoys state-of-theart performance on Sintel clean and final pass compared to published approaches with comparable computation and memory footprint. Codes and models will be released in https://github.com/DQiaole/MatchFlow.
****
Frustratingly Easy Regularization on Representation Can Boost Deep Reinforcement Learning
Qiang He, Huangyuan Su, Jieyu Zhang, Xinwen Hou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20215-20225
Deep reinforcement learning (DRL) gives the promise that an agent learns good policy from high-dimensional information, whereas representation learning removes irrelevant and redundant information and retains pertinent information. In this work, we demonstrate that the learned representation of the Q-network and its target Q-network should, in theory, satisfy a favorable distinguishable representation property. Specifically, there exists an upper bound on the representation similarity of the value functions of two adjacent time steps in a typical DRL setting. However, through illustrative experiments, we show that the learned DRL agent may violate this property and lead to a sub-optimal policy. Therefore, we propose a simple yet effective regularizer called Policy Evaluation with Easy Regularization on Representation (PEER), which aims to maintain the distinguishable representation property via explicit regularization on internal representations. And we provide the convergence rate guarantee of PEER. Implementing PEER requires only one line of code. Our experiments demonstrate that incorporating PEER into DRL can significantly improve performance and sample efficiency. Comprehensive experiments show that PEER achieves state-of-the-art performance on all 4 environments on PyBullet, 9 out of 12 tasks on DMControl, and 19 out of 26 games on Atari. To the best of our knowledge, PEER is the first work to study the inherent representation property of Q-network and its target. Our code is available at https://sites.google.com/view/peer-cvpr2023/.
****
PointDistiller: Structured Knowledge Distillation Towards Efficient and Compact 3D Detection
Linfeng Zhang, Runpei Dong, Hung-Shuo Tai, Kaisheng Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21791-21801
The remarkable breakthroughs in point cloud representation learning have boosted their usage in real-world applications such as self-driving cars and virtual reality. However, these applications usually have an urgent requirement for not only accurate but also efficient 3D object detection. Recently, knowledge distillation has been proposed as an effective model compression technique, which transfers the knowledge from an over-parameterized teacher to a lightweight student and achieves consistent effectiveness in 2D vision. However, due to point clouds' sparsity and irregularity, directly applying previous image-based knowledge distillation methods to point cloud detectors usually leads to unsatisfactory performance. To fill the gap, this paper proposes PointDistiller, a structured knowledge distillation framework for point clouds-based 3D detection. Concretely, PointDistiller includes local distillation which extracts and distills the local geom

etric structure of point clouds with dynamic graph convolution and reweighted le arning strategy, which highlights student learning on the critical points or vox els to improve knowledge distillation efficiency. Extensive experiments on both voxels-based and raw points-based detectors have demonstrated the effectiveness of our method over seven previous knowledge distillation methods. For instance, our 4X compressed PointPillars student achieves 2.8 and 3.4 mAP improvements on BEV and 3D object detection, outperforming its teacher by 0.9 and 1.8 mAP, respe ctively. Codes are available in the supplementary material and will be released on Github.
****
Learning Optical Expansion From Scale Matching
Han Ling, Yinghui Sun, Quansen Sun, Zhenwen Ren; Proceedings of the IEEE/CVF Con ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5445-5454
This paper address the problem of optical expansion (OE). OE describes the objec t scale change between two frames, widely used in monocular 3D vision tasks. Pre vious methods estimate optical expansion mainly from optical flow results, but t his two-stage architecture makes their results limited by the accuracy of optica l flow and less robust. To solve these problems, we propose the concept of 3D op tical flow by integrating optical expansion into the 2D optical flow, which is i mplemented by a plug-and-play module, namely TPCV. TPCV implements matching feat ures at the correct location and scale, thus allowing the simultaneous optimizat ion of optical flow and optical expansion tasks. Experimentally, we apply TPCV t o the RAFT optical flow baseline. Experimental results show that the baseline op tical flow performance is substantially improved. Moreover, we apply the optical flow and optical expansion results to various dynamic 3D vision tasks, includin g motion-in-depth, time-to-collision, and scene flow, often achieving significan t improvement over the prior SOTA. Code will be available at https://github.com/ HanLingsgjk/TPCV.
****
LEMaRT: Label-Efficient Masked Region Transform for Image Harmonization
Sheng Liu, Cong Phuoc Huynh, Cong Chen, Maxim Arap, Raffay Hamid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023 , pp. 18290-18299
We present a simple yet effective self-supervised pretraining method for image h armonization which can leverage large-scale unannotated image datasets. To achie ve this goal, we first generate pre-training data online with our Label-Efficien t Masked Region Transform (LEMaRT) pipeline. Given an image, LEMaRT generates a foreground mask and then applies a set of transformations to perturb various vis ual attributes, e.g., defocus blur, contrast, saturation, of the region specifie d by the generated mask. We then pre-train image harmonization models by recover ing the original image from the perturbed image. Secondly, we introduce an image harmonization model, namely SwinIH, by retrofitting the Swin Transformer [27] w ith a combination of local and global self-attention mechanisms. Pretraining Swi nIH with LEMaRT results in a new state of the art for image harmonization, while being label-efficient, i.e., consuming less annotated data for fine-tuning than existing methods. Notably, on iHarmony4 dataset [8], SwinIH outperforms the sta te of the art, i.e., SCS-Co [16] by a margin of 0.4 dB when it is fine-tuned on only 50% of the training data, and by 1.0 dB when it is trained on the full trai ning dataset.
****
How To Prevent the Poor Performance Clients for Personalized Federated Learning?
Zhe Qu, Xingyu Li, Xiao Han, Rui Duan, Chengchao Shen, Lixing Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 20 23, pp. 12167-12176
Personalized federated learning (pFL) collaboratively trains personalized models , which provides a customized model solution for individual clients in the prese nce of heterogeneous distributed local data. Although many recent studies have a pplied various algorithms to enhance personalization in pFL, they mainly focus o n improving the performance from averaging or top perspective. However, part of the clients may fall into poor performance and are not clearly discussed. Theref

ore, how to prevent these poor clients should be considered critically. Intuitiv ely, these poor clients may come from biased universal information shared with o thers. To address this issue, we propose a novel pFL strategy, called Personaliz e Locally, Generalize Universally (PLGU). PLGU generalizes the fine-grained univ ersal information and moderates its biased performance by designing a Layer-Wise d Sharpness Aware Minimization (LWSAM) algorithm while keeping the personalizati on local. Specifically, we embed our proposed PLGU strategy into two pFL schemes concluded in this paper: with/without a global model, and present the training procedures in detail. Through in-depth study, we show that the proposed PLGU str ategy achieves competitive generalization bounds on both considered pFL schemes. Our extensive experimental results show that all the proposed PLGU based-algori thms achieve state-of-the-art performance.
****
TopDiG: Class-Agnostic Topological Directional Graph Extraction From Remote Sens ing Images
Bingnan Yang, Mi Zhang, Zhan Zhang, Zhili Zhang, Xiangyun Hu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp . 1265-1274
Rapid development in automatic vector extraction from remote sensing images has been witnessed in recent years. However, the vast majority of existing works con centrate on a specific target, fragile to category variety, and hardly achieve s table performance crossing different categories. In this work, we propose an inn ovative class-agnostic model, namely TopDiG, to directly extract topological dir ectional graphs from remote sensing images and solve these issues. Firstly, TopD iG employs a topology-concentrated node detector (TCND) to detect nodes and obta in compact perception of topological components. Secondly, we propose a dynamic graph supervision (DGS) strategy to dynamically generate adjacency graph labels from unordered nodes. Finally, the directional graph (DiG) generator module is d esigned to construct topological directional graphs from predicted nodes. Experi ments on the Inria, CrowdAI, GID, GF2 and Massachusetts datasets empirically dem onstrate that TopDiG is class-agnostic and achieves competitive performance on a ll datasets.
****
Galactic: Scaling End-to-End Reinforcement Learning for Rearrangement at 100k St eps-per-Second
Vincent-Pierre Berges, Andrew Szot, Devendra Singh Chaplot, Aaron Gokaslan, Rooz beh Mottaghi, Dhruv Batra, Eric Undersander; Proceedings of the IEEE/CVF Confere nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13767-13777
We present Galactic, a large-scale simulation and reinforcement-learning (RL) fr amework for robotic mobile manipulation in indoor environments. Specifically, a Fetch robot (equipped with a mobile base, 7DoF arm, RGBD camera, egomotion, and onboard sensing) is spawned in a home environment and asked to rearrange objects -- by navigating to an object, picking it up, navigating to a target location, and then placing the object at the target location. Galactic is fast. In terms o f simulation speed (rendering + physics), Galactic achieves over 421,000 steps-p er-second (SPS) on an 8-GPU node, which is 54x faster than Habitat 2.0 (7699 SPS). More importantly, Galactic was designed to optimize the entire rendering+phys ics+RL interplay since any bottleneck in the interplay slows down training. In t erms of simulation+RL speed (rendering + physics + inference + learning), Galact ic achieves over 108,000 SPS, which 88x faster than Habitat 2.0 (1243 SPS). Thes e massive speed-ups not only drastically cut the wall-clock training time of exi sting experiments, but also unlock an unprecedented scale of new experiments. Fi rst, Galactic can train a mobile pick skill to >80% accuracy in under 16 minutes , a 100x speedup compared to the over 24 hours it takes to train the same skill in Habitat 2.0. Second, we use Galactic to perform the largest-scale experiment to date for rearrangement using 5B steps of experience in 46 hours, which is equ ivalent to 20 years of robot experience. This scaling results in a single neural network composed of task-agnostic components achieving 85% success in Geometric Goal rearrangement, compared to 0% success reported in Habitat 2.0 for the same approach. The code is available at github.com/facebookresearch/galactic.

****
StyleIPSB: Identity-Preserving Semantic Basis of StyleGAN for High Fidelity Face
 Swapping
Diqiong Jiang, Dan Song, Ruofeng Tong, Min Tang; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 352-361
Recent researches reveal that StyleGAN can generate highly realistic images, ins
piring researchers to use pretrained StyleGAN to generate high-fidelity swapped
faces. However, existing methods fail to meet the expectations in two essential
aspects of high-fidelity face swapping. Their results are blurry without pore-le
vel details and fail to preserve identity for challenging cases. To overcome the
 above artifacts, we innovatively construct a series of identity-preserving sema
ntic bases of StyleGAN (called StyleIPSB) in respect of pose, expression, and il
lumination. Each basis of StyleIPSB controls one specific semantic attribute and
 disentangles with the others. The StyleIPSB constrains style code in the subspa
ce of W+ space to preserve pore-level details. StyleIPSB gives us a novel tool f
or high-fidelity face swapping, and we propose a three-stage framework for face
swapping with StyleIPSB. Firstly, we transform the target facial images' attribu
tes to the source image. We learn the mapping from 3D Morphable Model (3DMM) par
ameters, which capture the prominent semantic variance, to the coordinates of St
yleIPSB that show higher identity-preserving and fidelity. Secondly, to transfor
m detailed attributes which 3DMM does not capture, we learn the residual attribu
te between the reenacted face and the target face. Finally, the face is blended
into the background of the target image. Extensive results and comparisons demon
strate that StyleIPSB can effectively preserve identity and pore-level details.
The results of face swapping can achieve state-of-the-art performance. We will r
elease our code at https://github.com/a686432/StyleIPSB.
****
Unknown Sniffer for Object Detection: Don't Turn a Blind Eye to Unknown Objects
Wenteng Liang, Feng Xue, Yihao Liu, Guofeng Zhong, Anlong Ming; Proceedings of t
he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 3230-3239
The recently proposed open-world object and open-set detection have achieved a b
reakthrough in finding never-seen-before objects and distinguishing them from kn
own ones. However, their studies on knowledge transfer from known classes to unk
nown ones are not deep enough, resulting in the scanty capability for detecting
unknowns hidden in the background. In this paper, we propose the unknown sniffer
 (UnSniffer) to find both unknown and known objects. Firstly, the generalized ob
ject confidence (GOC) score is introduced, which only uses known samples for sup
ervision and avoids improper suppression of unknowns in the background. Signific
antly, such confidence score learned from known objects can be generalized to un
known ones. Additionally, we propose a negative energy suppression loss to furth
er suppress the non-object samples in the background. Next, the best box of each
 unknown is hard to obtain during inference due to lacking their semantic inform
ation in training. To solve this issue, we introduce a graph-based determination
 scheme to replace hand-designed non-maximum suppression (NMS) post-processing.
Finally, we present the Unknown Object Detection Benchmark, the first publicly b
enchmark that encompasses precision evaluation for unknown detection to our know
ledge. Experiments show that our method is far better than the existing state-of
-the-art methods. Code is available at: https://github.com/Went-Liang/UnSniffer.
****
Discriminator-Cooperated Feature Map Distillation for GAN Compression
Tie Hu, Mingbao Lin, Lizhou You, Fei Chao, Rongrong Ji; Proceedings of the IEEE/
CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2035
1-20360
Despite excellent performance in image generation, Generative Adversarial Networ
ks (GANs) are notorious for its requirements of enormous storage and intensive c
omputation. As an awesome "performance maker", knowledge distillation is demonst
rated to be particularly efficacious in exploring low-priced GANs. In this paper
, we investigate the irreplaceability of teacher discriminator and present an in
ventive discriminator-cooperated distillation, abbreviated as DCD, towards refin

ing better feature maps from the generator. In contrast to conventional pixel-to-pixel match methods in feature map distillation, our DCD utilizes teacher discriminator as a transformation to drive intermediate results of the student generator to be perceptually close to corresponding outputs of the teacher generator. Furthermore, in order to mitigate mode collapse in GAN compression, we construct a collaborative adversarial training paradigm where the teacher discriminator is from scratch established to co-train with student generator in company with our DCD. Our DCD shows superior results compared with existing GAN compression methods. For instance, after reducing over 40x MACs and 80x parameters of CycleGAN, we well decrease FID metric from 61.53 to 48.24 while the current SoTA method merely has 51.92. This work's source code has been made accessible at https://github.com/poopit/DCD-official.
****
Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection
Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Yunchao Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12105-12114
Recently, there has been a significant advancement in image generation technology, known as GAN. It can easily generate realistic fake images, leading to an increased risk of abuse. However, most image detectors suffer from sharp performance drops in unseen domains. The key of fake image detection is to develop a generalized representation to describe the artifacts produced by generation models. In this work, we introduce a novel detection framework, named Learning on Gradients (LGrad), designed for identifying GAN-generated images, with the aim of constructing a generalized detector with cross-model and cross-data. Specifically, a pretrained CNN model is employed as a transformation model to convert images into gradients. Subsequently, we leverage these gradients to present the generalized artifacts, which are fed into the classifier to ascertain the authenticity of the images. In our framework, we turn the data-dependent problem into a transformation-model-dependent problem. To the best of our knowledge, this is the first study to utilize gradients as the representation of artifacts in GAN-generated images. Extensive experiments demonstrate the effectiveness and robustness of gradients as generalized artifact representations. Our detector achieves a new state-of-the-art performance with a remarkable gain of 11.4%. The code is released at https://github.com/chuangchuangtan/LGrad.
****
Don't Lie to Me! Robust and Efficient Explainability With Verified Perturbation Analysis
Thomas Fel, Melanie Ducoffe, David Vigouroux, Rémi Cadène, Mikaël Capelle, Claire Nicodème, Thomas Serre; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16153-16163
A variety of methods have been proposed to try to explain how deep neural networks make their decisions. Key to those approaches is the need to sample the pixel space efficiently in order to derive importance maps. However, it has been shown that the sampling methods used to date introduce biases and other artifacts, leading to inaccurate estimates of the importance of individual pixels and severely limit the reliability of current explainability methods. Unfortunately, the alternative -- to exhaustively sample the image space is computationally prohibitive. In this paper, we introduce EVA (Explaining using Verified perturbation Analysis) -- the first explainability method guarantee to have an exhaustive exploration of a perturbation space. Specifically, we leverage the beneficial properties of verified perturbation analysis -- time efficiency, tractability and guaranteed complete coverage of a manifold -- to efficiently characterize the input variables that are most likely to drive the model decision. We evaluate the approach systematically and demonstrate state-of-the-art results on multiple benchmarks.
****
StyleAdv: Meta Style Adversarial Training for Cross-Domain Few-Shot Learning
Yuqian Fu, Yu Xie, Yanwei Fu, Yu-Gang Jiang; Proceedings of the IEEE/CVF Confere

nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24575-24584
Cross-Domain Few-Shot Learning (CD-FSL) is a recently emerging task that tackles few-shot learning across different domains. It aims at transferring prior knowledge learned on the source dataset to novel target datasets. The CD-FSL task is especially challenged by the huge domain gap between different datasets. Critically, such a domain gap actually comes from the changes of visual styles, and wave-SAN empirically shows that spanning the style distribution of the source data helps alleviate this issue. However, wave-SAN simply swaps styles of two images. Such a vanilla operation makes the generated styles "real" and "easy", which still fall into the original set of the source styles. Thus, inspired by vanilla adversarial learning, a novel model-agnostic meta Style Adversarial training (StyleAdv) method together with a novel style adversarial attack method is proposed for CD-FSL. Particularly, our style attack method synthesizes both "virtual" and "hard" adversarial styles for model training. This is achieved by perturbing the original style with the signed style gradients. By continually attacking styles and forcing the model to recognize these challenging adversarial styles, our model is gradually robust to the visual styles, thus boosting the generalization ability for novel target datasets. Besides the typical CNN-based backbone, we also employ our StyleAdv method on large-scale pretrained vision transformer. Extensive experiments conducted on eight various target datasets show the effectiveness of our method. Whether built upon ResNet or ViT, we achieve the new state of the art for CD-FSL. Code is available at https://github.com/lovelyqian/StyleAdv-CDFSL.
****
Multi-Concept Customization of Text-to-Image Diffusion
Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, Jun-Yan Zhu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1931-1941
While generative models produce high-quality images of concepts learned from a large-scale database, a user often wishes to synthesize instantiations of their own concepts (for example, their family, pets, or items). Can we teach a model to quickly acquire a new concept, given a few examples? Furthermore, can we compose multiple new concepts together? We propose Custom Diffusion, an efficient method for augmenting existing text-to-image models. We find that only optimizing a few parameters in the text-to-image conditioning mechanism is sufficiently powerful to represent new concepts while enabling fast tuning (6 minutes). Additionally, we can jointly train for multiple concepts or combine multiple fine-tuned models into one via closed-form constrained optimization. Our fine-tuned model generates variations of multiple new concepts and seamlessly composes them with existing concepts in novel settings. Our method outperforms or performs on par with several baselines and concurrent works in both qualitative and quantitative evaluations, while being memory and computationally efficient.
****
Defending Against Patch-Based Backdoor Attacks on Self-Supervised Learning
Ajinkya Tejankar, Maziar Sanjabi, Qifan Wang, Sinong Wang, Hamed Firooz, Hamed Pirsiavash, Liang Tan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12239-12249
Recently, self-supervised learning (SSL) was shown to be vulnerable to patch-based data poisoning backdoor attacks. It was shown that an adversary can poison a small part of the unlabeled data so that when a victim trains an SSL model on it, the final model will have a backdoor that the adversary can exploit. This work aims to defend self-supervised learning against such attacks. We use a three-step defense pipeline, where we first train a model on the poisoned data. In the second step, our proposed defense algorithm (PatchSearch) uses the trained model to search the training data for poisoned samples and removes them from the training set. In the third step, a final model is trained on the cleaned-up training set. Our results show that PatchSearch is an effective defense. As an example, it improves a model's accuracy on images containing the trigger from 38.2% to 63.7% which is very close to the clean model's accuracy, 64.6%. Moreover, we show that PatchSearch outperforms baselines and state-of-the-art defense approaches in

cluding those using additional clean, trusted data. Our code is available at htt
ps://github.com/UCDvision/PatchSearch
****
Long-Tailed Visual Recognition via Self-Heterogeneous Integration With Knowledge
 Excavation
Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, Hanzi Wang; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23
695-23704
Deep neural networks have made huge progress in the last few decades. However, a
s the real-world data often exhibits a long-tailed distribution, vanilla deep mo
dels tend to be heavily biased toward the majority classes. To address this prob
lem, state-of-the-art methods usually adopt a mixture of experts (MoE) to focus
on different parts of the long-tailed distribution. Experts in these methods are
 with the same model depth, which neglects the fact that different classes may h
ave different preferences to be fit by models with different depths. To this end
, we propose a novel MoE-based method called Self-Heterogeneous Integration with
 Knowledge Excavation (SHIKE). We first propose Depth-wise Knowledge Fusion (DKF
) to fuse features between different shallow parts and the deep part in one netw
ork for each expert, which makes experts more diverse in terms of representation
. Based on DKF, we further propose Dynamic Knowledge Transfer (DKT) to reduce th
e influence of the hardest negative class that has a non-negligible impact on th
e tail classes in our MoE framework. As a result, the classification accuracy of
 long-tailed data can be significantly improved, especially for the tail classes
. SHIKE achieves the state-of-the-art performance of 56.3%, 60.3%, 75.4%, and 41
.9% on CIFAR100-LT (IF100), ImageNet-LT, iNaturalist 2018, and Places-LT, respec
tively. The source code is available at https://github.com/jinyan-06/SHIKE.
****
GeoNet: Benchmarking Unsupervised Adaptation Across Geographies
Tarun Kalluri, Wangdong Xu, Manmohan Chandraker; Proceedings of the IEEE/CVF Con
ference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15368-15379
In recent years, several efforts have been aimed at improving the robustness of
vision models to domains and environments unseen during training. An important p
ractical problem pertains to models deployed in a new geography that is under-re
presented in the training dataset, posing a direct challenge to fair and inclusi
ve computer vision. In this paper, we study the problem of geographic robustness
 and make three main contributions. First, we introduce a large-scale dataset Ge
oNet for geographic adaptation containing benchmarks across diverse tasks like s
cene recognition (GeoPlaces), image classification (GeoImNet) and universal adap
tation (GeoUniDA). Second, we investigate the nature of distribution shifts typi
cal to the problem of geographic adaptation and hypothesize that the major sourc
e of domain shifts arise from significant variations in scene context (context s
hift), object design (design shift) and label distribution (prior shift) across
geographies. Third, we conduct an extensive evaluation of several state-of-the-a
rt unsupervised domain adaptation algorithms and architectures on GeoNet, showin
g that they do not suffice for geographical adaptation, and that large-scale pre
-training using large vision models also does not lead to geographic robustness.
 Our dataset is publicly available at https://tarun005.github.io/GeoNet.
****
Context De-Confounded Emotion Recognition
Dingkang Yang, Zhaoyu Chen, Yuzheng Wang, Shunli Wang, Mingcheng Li, Siao Liu, X
iao Zhao, Shuai Huang, Zhiyan Dong, Peng Zhai, Lihua Zhang; Proceedings of the I
EEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp.
19005-19015
Context-Aware Emotion Recognition (CAER) is a crucial and challenging task that
aims to perceive the emotional states of the target person with contextual infor
mation. Recent approaches invariably focus on designing sophisticated architectu
res or mechanisms to extract seemingly meaningful representations from subjects
and contexts. However, a long-overlooked issue is that a context bias in existin
g datasets leads to a significantly unbalanced distribution of emotional states
among different context scenarios. Concretely, the harmful bias is a confounder

that misleads existing models to learn spurious correlations based on conventional likelihood estimation, significantly limiting the models' performance. To tackle the issue, this paper provides a causality-based perspective to disentangle the models from the impact of such bias, and formulate the causalities among variables in the CAER task via a tailored causal graph. Then, we propose a Contextual Causal Intervention Module (CCIM) based on the backdoor adjustment to de-confound the confounder and exploit the true causal effect for model training. CCIM is plug-in and model-agnostic, which improves diverse state-of-the-art approaches by considerable margins. Extensive experiments on three benchmark datasets demonstrate the effectiveness of our CCIM and the significance of causal insight.
****
LinK: Linear Kernel for LiDAR-Based 3D Perception
Tao Lu, Xiang Ding, Haisong Liu, Gangshan Wu, Limin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1105-1115
Extending the success of 2D Large Kernel to 3D perception is challenging due to: 1. the cubically-increasing overhead in processing 3D data; 2. the optimization difficulties from data scarcity and sparsity. Previous work has taken the first step to scale up the kernel size from 3x3x3 to 7x7x7 by introducing block-shared weights. However, to reduce the feature variations within a block, it only employs modest block size and fails to achieve larger kernels like the 21x21x21. To address this issue, we propose a new method, called LinK, to achieve a wider-range perception receptive field in a convolution-like manner with two core designs. The first is to replace the static kernel matrix with a linear kernel generator, which adaptively provides weights only for non-empty voxels. The second is to reuse the pre-computed aggregation results in the overlapped blocks to reduce computation complexity. The proposed method successfully enables each voxel to perceive context within a range of 21x21x21. Extensive experiments on two basic perception tasks, 3D object detection and 3D semantic segmentation, demonstrate the effectiveness of our method. Notably, we rank 1st on the public leaderboard of the 3D detection benchmark of nuScenes (LiDAR track), by simply incorporating a LinK-based backbone into the basic detector, CenterPoint. We also boost the strong segmentation baseline's mIoU with 2.7% in the SemanticKITTI test set. Code is available at https://github.com/MCG-NJU/LinK.
****
CP3: Channel Pruning Plug-In for Point-Based Networks
Yaomin Huang, Ning Liu, Zhengping Che, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Guixu Zhang, Xinmei Liu, Feifei Feng, Jian Tang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5302-5312
Channel pruning has been widely studied as a prevailing method that effectively reduces both computational cost and memory footprint of the original network while keeping a comparable accuracy performance. Though great success has been achieved in channel pruning for 2D image-based convolutional networks (CNNs), existing works seldom extend the channel pruning methods to 3D point-based neural networks (PNNs). Directly implementing the 2D CNN channel pruning methods to PNNs undermine the performance of PNNs because of the different representations of 2D images and 3D point clouds as well as the network architecture disparity. In this paper, we proposed CP^3, which is a Channel Pruning Plug-in for Point-based network. CP^3 is elaborately designed to leverage the characteristics of point clouds and PNNs in order to enable 2D channel pruning methods for PNNs. Specifically, it presents a coordinate-enhanced channel importance metric to reflect the correlation between dimensional information and individual channel features, and it recycles the discarded points in PNN's sampling process and reconsiders their potentially-exclusive information to enhance the robustness of channel pruning. Experiments on various PNN architectures show that CP^3 constantly improves state-of-the-art 2D CNN pruning approaches on different point cloud tasks. For instance, our compressed PointNeXt-S on ScanObjectNN achieves an accuracy of 88.52% with a pruning rate of 57.8%, outperforming the baseline pruning methods with an accuracy gain of 1.94%.
****

InstructPix2Pix: Learning To Follow Image Editing Instructions
Tim Brooks, Aleksander Holynski, Alexei A. Efros; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18392-18402
We propose a method for editing images from human instructions: given an input image and a written instruction that tells the model what to do, our model follows these instructions to edit the image. To obtain training data for this problem, we combine the knowledge of two large pretrained models--a language model (GPT-3) and a text-to-image model (Stable Diffusion)--to generate a large dataset of image editing examples. Our conditional diffusion model, InstructPix2Pix, is trained on our generated data, and generalizes to real images and user-written instructions at inference time. Since it performs edits in the forward pass and does not require per-example fine-tuning or inversion, our model edits images quickly, in a matter of seconds. We show compelling editing results for a diverse collection of input images and written instructions.
****
Learning Transformation-Predictive Representations for Detection and Description of Local Features
Zihao Wang, Chunxu Wu, Yifei Yang, Zhen Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11464-11473
The task of key-points detection and description is to estimate the stable location and discriminative representation of local features, which is essential for image matching. However, either the rough hard positive or negative labels generated from one-to-one correspondences among images bring indistinguishable samples, called pseudo positives or negatives, which act as inconsistent supervisions while learning key-points used for matching. Such pseudo-labeled samples prevent deep neural networks from learning discriminative descriptions for accurate matching. To tackle this challenge, we propose to learn transformation-predictive representations with self-supervised contrastive learning. We maximize the similarity between corresponded views of the same 3D point (landmark) by using none of the negative sample pairs (including true and pseudo negatives) and avoiding collapsing solutions. Then we design a learnable label prediction mechanism to soften the hard positive labels into soft continuous targets. The aggressively updated soft labels extensively deal with the training bottleneck (derived from the label noise of pseudo positives) and make the model can be trained under a stronger augmentation paradigm. Our self-supervised method outperforms the state-of-the-art on the standard image matching benchmarks by noticeable margins and shows excellent generalization capability on multiple downstream tasks.
****
Two-Way Multi-Label Loss
Takumi Kobayashi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7476-7485
A natural image frequently contains multiple classification targets, accordingly providing multiple class labels rather than a single label per image. While the single-label classification is effectively addressed by applying a softmax cross-entropy loss, the multi-label task is tackled mainly in a binary cross-entropy (BCE) framework. In contrast to the softmax loss, the BCE loss involves issues regarding imbalance as multiple classes are decomposed into a bunch of binary classifications; recent works improve the BCE loss to cope with the issue by means of weighting. In this paper, we propose a multi-label loss by bridging a gap between the softmax loss and the multi-label scenario. The proposed loss function is formulated on the basis of relative comparison among classes which also enables us to further improve discriminative power of features by enhancing classification margin. The loss function is so flexible as to be applicable to a multi-label setting in two ways for discriminating classes as well as samples. In the experiments on multi-label classification, the proposed method exhibits competitive performance to the other multi-label losses, and it also provides transferrable features on single-label ImageNet training. Codes are available at https://github.com/tk1980/TwowayMultiLabelLoss.
****

Progressive Disentangled Representation Learning for Fine-Grained Controllable T
alking Head Synthesis
Duomin Wang, Yu Deng, Zixin Yin, Heung-Yeung Shum, Baoyuan Wang; Proceedings of
the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023,
pp. 17979-17989
We present a novel one-shot talking head synthesis method that achieves disentan
gled and fine-grained control over lip motion, eye gaze&blink, head pose, and em
otional expression. We represent different motions via disentangled latent repre
sentations and leverage an image generator to synthesize talking heads from them
. To effectively disentangle each motion factor, we propose a progressive disent
angled representation learning strategy by separating the factors in a coarse-to
-fine manner, where we first extract unified motion feature from the driving sig
nal, and then isolate each fine-grained motion from the unified feature. We intr
oduce motion-specific contrastive learning and regressing for non-emotional moti
ons, and feature-level decorrelation and self-reconstruction for emotional expre
ssion, to fully utilize the inherent properties of each motion factor in unstruc
tured video data to achieve disentanglement. Experiments show that our method pr
ovides high quality speech&lip-motion synchronization along with precise and dis
entangled control over multiple extra facial motions, which can hardly be achiev
ed by previous methods.
****
Breaking the "Object" in Video Object Segmentation
Pavel Tokmakov, Jie Li, Adrien Gaidon; Proceedings of the IEEE/CVF Conference on
Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22836-22845
The appearance of an object can be fleeting when it transforms. As eggs are brok
en or paper is torn, their color, shape, and texture can change dramatically, pr
eserving virtually nothing of the original except for the identity itself. Yet,
this important phenomenon is largely absent from existing video object segmentat
ion (VOS) benchmarks. In this work, we close the gap by collecting a new dataset
for Video Object Segmentation under Transformations (VOST). It consists of more
than 700 high-resolution videos, captured in diverse environments, which are 20
seconds long on average and densely labeled with instance masks. A careful, mul
ti-step approach is adopted to ensure that these videos focus on complex object
transformations, capturing their full temporal extent. We then extensively evalu
ate state-of-the-art VOS methods and make a number of important discoveries. In
particular, we show that existing methods struggle when applied to this novel ta
sk and that their main limitation lies in over-reliance on static, appearance cu
es. This motivates us to propose a few modifications for the top-performing base
line that improve its performance by better capturing spatio-temporal informatio
n. But more broadly, the hope is to stimulate discussion on learning more robust
video object representations.
****
Where Is My Wallet? Modeling Object Proposal Sets for Egocentric Visual Query Lo
calization
Mengmeng Xu, Yanghao Li, Cheng-Yang Fu, Bernard Ghanem, Tao Xiang, Juan-Manuel P
érez-Rúa; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR), 2023, pp. 2593-2603
This paper deals with the problem of localizing objects in image and video datas
ets from visual exemplars. In particular, we focus on the challenging problem of
egocentric visual query localization. We first identify grave implicit biases i
n current query-conditioned model design and visual query datasets. Then, we dir
ectly tackle such biases at both frame and object set levels. Concretely, our me
thod solves these issues by expanding limited annotations and dynamically droppi
ng object proposals during training. Additionally, we propose a novel transforme
r-based module that allows for object-proposal set context to be considered whil
e incorporating query information. We name our module Conditioned Contextual Tra
nsformer or CocoFormer. Our experiments show that the proposed adaptations impro
ve egocentric query detection, leading to a better visual query localization sys
tem in both 2D and 3D configurations. Thus, we are able to improve frame-level d
etection performance from 26.28% to 31.26% in AP, which correspondingly improves

the VQ2D and VQ3D localization scores by significant margins. Our improved cont ext-aware query object detector ranked first and second in the VQ2D and VQ3D tas ks in the 2nd Ego4D challenge. In addition, we showcase the relevance of our pro posed model in the Few-Shot Detection (FSD) task, where we also achieve SOTA res ults.
****
Dionysus: Recovering Scene Structures by Dividing Into Semantic Pieces
Likang Wang, Lei Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12576-12587
Most existing 3D reconstruction methods result in either detail loss or unsatisf ying efficiency. However, effectiveness and efficiency are equally crucial in re al-world applications, e.g., autonomous driving and augmented reality. We argue that this dilemma comes from wasted resources on valueless depth samples. This p aper tackles the problem by proposing a novel learning-based 3D reconstruction f ramework named Dionysus. Our main contribution is to find out the most promising depth candidates from estimated semantic maps. This strategy simultaneously ena bles high effectiveness and efficiency by attending to the most reliable nominat ors. Specifically, we distinguish unreliable depth candidates by checking the cr oss-view semantic consistency and allow adaptive sampling by redistributing dept h nominators among pixels. Experiments on the most popular datasets confirm our proposed framework's effectiveness.
****
ReDirTrans: Latent-to-Latent Translation for Gaze and Head Redirection
Shiwei Jin, Zhen Wang, Lei Wang, Ning Bi, Truong Nguyen; Proceedings of the IEEE /CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 554 7-5556
Learning-based gaze estimation methods require large amounts of training data wi th accurate gaze annotations. Facing such demanding requirements of gaze data co llection and annotation, several image synthesis methods were proposed, which su ccessfully redirected gaze directions precisely given the assigned conditions. H owever, these methods focused on changing gaze directions of the images that onl y include eyes or restricted ranges of faces with low resolution (less than 128* 128) to largely reduce interference from other attributes such as hairs, which l imits application scenarios. To cope with this limitation, we proposed a portabl e network, called ReDirTrans, achieving latent-to-latent translation for redirec ting gaze directions and head orientations in an interpretable manner. ReDirTran s projects input latent vectors into aimed-attribute embeddings only and redirec ts these embeddings with assigned pitch and yaw values. Then both the initial an d edited embeddings are projected back (deprojected) to the initial latent space as residuals to modify the input latent vectors by subtraction and addition, re presenting old status removal and new status addition. The projection of aimed a ttributes only and subtraction-addition operations for status replacement essent ially mitigate impacts on other attributes and the distribution of latent vector s. Thus, by combining ReDirTrans with a pretrained fixed e4e-StyleGAN pair, we c reated ReDirTrans-GAN, which enables accurately redirecting gaze in full-face im ages with 1024*1024 resolution while preserving other attributes such as identit y, expression, and hairstyle. Furthermore, we presented improvements for the dow nstream learning-based gaze estimation task, using redirected samples as dataset augmentation.
****
Advancing Visual Grounding With Scene Knowledge: Benchmark and Method
Zhihong Chen, Ruifei Zhang,Yibing Song, Xiang Wan, Guanbin Li; Proceedings of th e IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, p p. 15039-15049
Visual grounding (VG) aims to establish fine-grained alignment between vision an d language. Ideally, it can be a testbed for vision-and-language models to evalu ate their understanding of the images and texts and their reasoning abilities ov er their joint space. However, most existing VG datasets are constructed using s imple description texts, which do not require sufficient reasoning over the imag es and texts. This has been demonstrated in a recent study, where a simple LSTM-

based text encoder without pretraining can achieve state-of-the-art performance on mainstream VG datasets. Therefore, in this paper, we propose a novel benchmar k of Scene Knowledge-guided Visual Grounding (SK-VG), where the image content an d referring expressions are not sufficient to ground the target objects, forcing the models to have a reasoning ability on the long-form scene knowledge. To per form this task, we propose two approaches to accept the triple-type input, where the former embeds knowledge into the image features before the image-query inte raction; the latter leverages linguistic structure to assist in computing the im age-text matching. We conduct extensive experiments to analyze the above methods and show that the proposed approaches achieve promising results but still leave room for improvement, including performance and interpretability.
****
Noisy Correspondence Learning With Meta Similarity Correction
Haochen Han, Kaiyao Miao, Qinghua Zheng, Minnan Luo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7517-75 26
Despite the success of multimodal learning in cross-modal retrieval task, the re markable progress relies on the correct correspondence among multimedia data. Ho wever, collecting such ideal data is expensive and time-consuming. In practice, most widely used datasets are harvested from the Internet and inevitably contain mismatched pairs. Training on such noisy correspondence datasets causes perform ance degradation because the cross-modal retrieval methods can wrongly enforce t he mismatched data to be similar. To tackle this problem, we propose a Meta Simi larity Correction Network (MSCN) to provide reliable similarity scores. We view a binary classification task as the meta-process that encourages the MSCN to lea rn discrimination from positive and negative meta-data. To further alleviate the influence of noise, we design an effective data purification strategy using met a-data as prior knowledge to remove the noisy samples. Extensive experiments are conducted to demonstrate the strengths of our method in both synthetic and real -world noises, including Flickr30K, MS-COCO, and Conceptual Captions.
****
CoWs on Pasture: Baselines and Benchmarks for Language-Driven Zero-Shot Object N avigation
Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, Shuran Song; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco gnition (CVPR), 2023, pp. 23171-23181
For robots to be generally useful, they must be able to find arbitrary objects d escribed by people (i.e., be language-driven) even without expensive navigation training on in-domain data (i.e., perform zero-shot inference). We explore these capabilities in a unified setting: language-driven zero-shot object navigation (L-ZSON). Inspired by the recent success of open-vocabulary models for image cla ssification, we investigate a straightforward framework, CLIP on Wheels (CoW), t o adapt open-vocabulary models to this task without fine-tuning. To better evalu ate L-ZSON, we introduce the Pasture benchmark, which considers finding uncommon objects, objects described by spatial and appearance attributes, and hidden obj ects described relative to visible objects. We conduct an in-depth empirical stu dy by directly deploying 22 CoW baselines across Habitat, RoboTHOR, and Pasture. In total we evaluate over 90k navigation episodes and find that (1) CoW baselin es often struggle to leverage language descriptions, but are surprisingly profic ient at finding uncommon objects. (2) A simple CoW, with CLIP-based object local ization and classical exploration---and no additional training---matches the nav igation efficiency of a state-of-the-art ZSON method trained for 500M steps on H abitat MP3D data. This same CoW provides a 15.6 percentage point improvement in success over a state-of-the-art RoboTHOR ZSON model.
****
CIGAR: Cross-Modality Graph Reasoning for Domain Adaptive Object Detection
Yabo Liu, Jinghua Wang, Chao Huang, Yaowei Wang, Yong Xu; Proceedings of the IEE E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23 776-23786
Unsupervised domain adaptive object detection (UDA-OD) aims to learn a detector

by generalizing knowledge from a labeled source domain to an unlabeled target do main. Though the existing graph-based methods for UDA-OD perform well in some ca ses, they cannot learn a proper node set for the graph. In addition, these metho ds build the graph solely based on the visual features and do not consider the l inguistic knowledge carried by the semantic prototypes, e.g., dataset labels. To overcome these problems, we propose a cross-modality graph reasoning adaptation (CIGAR) method to take advantage of both visual and linguistic knowledge. Speci fically, our method performs cross-modality graph reasoning between the linguist ic modality graph and visual modality graphs to enhance their representations. W e also propose a discriminative feature selector to find the most discriminative features and take them as the nodes of the visual graph for both efficiency and effectiveness. In addition, we employ the linguistic graph matching loss to reg ulate the update of linguistic graphs and maintain their semantic representation during the training process. Comprehensive experiments validate the effectivene ss of our proposed CIGAR.
****
Multiview Compressive Coding for 3D Reconstruction
Chao-Yuan Wu, Justin Johnson, Jitendra Malik, Christoph Feichtenhofer, Georgia Gkioxari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9065-9075
A central goal of visual recognition is to understand objects and scenes from a single image. 2D recognition has witnessed tremendous progress thanks to large-s cale learning and general-purpose representations. But, 3D poses new challenges stemming from occlusions not depicted in the image. Prior works try to overcome these by inferring from multiple views or rely on scarce CAD models and category -specific priors which hinder scaling to novel settings. In this work, we explor e single-view 3D reconstruction by learning generalizable representations inspir ed by advances in self-supervised learning. We introduce a simple framework that operates on 3D points of single objects or whole scenes coupled with category-a gnostic large-scale training from diverse RGB-D videos. Our model, Multiview Com pressive Coding (MCC), learns to compress the input appearance and geometry to p redict the 3D structure by querying a 3D-aware decoder. MCC's generality and eff iciency allow it to learn from large-scale and diverse data sources with strong generalization to novel objects imagined by DALL*E 2 or captured in-the-wild wit h an iPhone.
****
HOOD: Hierarchical Graphs for Generalized Modelling of Clothing Dynamics
Artur Grigorev, Michael J. Black, Otmar Hilliges; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16965-16974
We propose a method that leverages graph neural networks, multi-level message pa ssing, and unsupervised training to enable real-time prediction of realistic clo thing dynamics. Whereas existing methods based on linear blend skinning must be trained for specific garments, our method is agnostic to body shape and applies to tight-fitting garments as well as loose, free-flowing clothing. Our method fu rthermore handles changes in topology (e.g., garments with buttons or zippers) a nd material properties at inference time. As one key contribution, we propose a hierarchical message-passing scheme that efficiently propagates stiff stretching modes while preserving local detail. We empirically show that our method outper forms strong baselines quantitatively and that its results are perceived as more realistic than state-of-the-art methods.
****
HyperReel: High-Fidelity 6-DoF Video With Ray-Conditioned Sampling
Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhöfer, Johannes Kopf, Matthew O’Toole, Changil Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16610-16620
Volumetric scene representations enable photorealistic view synthesis for static scenes and form the basis of several existing 6-DoF video techniques. However, the volume rendering procedures that drive these representations necessitate car eful trade-offs in terms of quality, rendering speed, and memory efficiency. In

particular, existing methods fail to simultaneously achieve real-time performanc e, small memory footprint, and high-quality rendering for challenging real-world scenes. To address these issues, we present HyperReel --- a novel 6-DoF video r epresentation. The two core components of HyperReel are: (1) a ray-conditioned s ample prediction network that enables high-fidelity, high frame rate rendering a t high resolutions and (2) a compact and memory-efficient dynamic volume represe ntation. Our 6-DoF video pipeline achieves the best performance compared to prio r and contemporary approaches in terms of visual quality with small memory requi rements, while also rendering at up to 18 frames-per-second at megapixel resolut ion without any custom CUDA code.
****
Rethinking Video ViTs: Sparse Video Tubes for Joint Image and Video Learning
AJ Piergiovanni, Weicheng Kuo, Anelia Angelova; Proceedings of the IEEE/CVF Conf erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2214-2224
We present a simple approach which can turn a ViT encoder into an efficient vide o model, which can seamlessly work with both image and video inputs. By sparsely sampling the inputs, the model is able to do training and inference from both i nputs. The model is easily scalable and can be adapted to large-scale pre-traine d ViTs without requiring full finetuning. The model achieves SOTA results.
****
Modeling Entities As Semantic Points for Visual Information Extraction in the Wi ld
Zhibo Yang, Rujiao Long, Pengfei Wang, Sibo Song, Humen Zhong, Wenqing Cheng, Xi ang Bai, Cong Yao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15358-15367
Recently, Visual Information Extraction (VIE) has been becoming increasingly imp ortant in both academia and industry, due to the wide range of real-world applic ations. Previously, numerous works have been proposed to tackle this problem. Ho wever, the benchmarks used to assess these methods are relatively plain, i.e., s cenarios with real-world complexity are not fully represented in these benchmark s. As the first contribution of this work, we curate and release a new dataset f or VIE, in which the document images are much more challenging in that they are taken from real applications, and difficulties such as blur, partial occlusion, and printing shift are quite common. All these factors may lead to failures in i nformation extraction. Therefore, as the second contribution, we explore an alte rnative approach to precisely and robustly extract key information from document images under such tough conditions. Specifically, in contrast to previous metho ds, which usually either incorporate visual information into a multi-modal archi tecture or train text spotting and information extraction in an end-to-end fashi on, we explicitly model entities as semantic points, i.e., center points of enti ties are enriched with semantic information describing the attributes and relati onships of different entities, which could largely benefit entity labeling and l inking. Extensive experiments on standard benchmarks in this field as well as th e proposed dataset demonstrate that the proposed method can achieve significantl y enhanced performance on entity labeling and linking, compared with previous st ate-of-the-art models.
****
MobileVOS: Real-Time Video Object Segmentation Contrastive Learning Meets Knowle dge Distillation
Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, Albert Saà-Garriga; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2 023, pp. 10480-10490
This paper tackles the problem of semi-supervised video object segmentation on r esource-constrained devices, such as mobile phones. We formulate this problem as a distillation task, whereby we demonstrate that small space-time-memory networ ks with finite memory can achieve competitive results with state of the art, but at a fraction of the computational cost (32 milliseconds per frame on a Samsung Galaxy S22). Specifically, we provide a theoretically grounded framework that u nifies knowledge distillation with supervised contrastive representation learnin g. These models are able to jointly benefit from both pixel-wise contrastive lea

rning and distillation from a pre-trained teacher. We validate this loss by achi eving competitive J&F to state of the art on both the standard DAVIS and YouTube benchmarks, despite running up to x5 faster, and with x32 fewer parameters.
****
PCR: Proxy-Based Contrastive Replay for Online Class-Incremental Continual Learn ing
Huiwei Lin, Baoquan Zhang, Shanshan Feng, Xutao Li, Yunming Ye; Proceedings of t he IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24246-24255
Online class-incremental continual learning is a specific task of continual lear ning. It aims to continuously learn new classes from data stream and the samples of data stream are seen only once, which suffers from the catastrophic forgetti ng issue, i.e., forgetting historical knowledge of old classes. Existing replay-based methods effectively alleviate this issue by saving and replaying part of o ld data in a proxy-based or contrastive-based replay manner. Although these two replay manners are effective, the former would incline to new classes due to cla ss imbalance issues, and the latter is unstable and hard to converge because of the limited number of samples. In this paper, we conduct a comprehensive analysi s of these two replay manners and find that they can be complementary. Inspired by this finding, we propose a novel replay-based method called proxy-based contr astive replay (PCR). The key operation is to replace the contrastive samples of anchors with corresponding proxies in the contrastive-based way. It alleviates t he phenomenon of catastrophic forgetting by effectively addressing the imbalance issue, as well as keeps a faster convergence of the model. We conduct extensive experiments on three real-world benchmark datasets, and empirical results consi stently demonstrate the superiority of PCR over various state-of-the-art methods .
****
Pose Synchronization Under Multiple Pair-Wise Relative Poses
Yifan Sun, Qixing Huang; Proceedings of the IEEE/CVF Conference on Computer Visi on and Pattern Recognition (CVPR), 2023, pp. 13072-13081
Pose synchronization, which seeks to estimate consistent absolute poses among a collection of objects from noisy relative poses estimated between pairs of objec ts in isolation, is a fundamental problem in many inverse applications. This pap er studies an extreme setting where multiple relative pose estimates exist betwe en each object pair, and the majority is incorrect. Popular methods that solve p ose synchronization via recovering a low-rank matrix that encodes relative poses in block fail under this extreme setting. We introduce a three-step algorithm f or pose synchronization under multiple relative pose inputs. The first step perf orms diffusion and clustering to compute the candidate poses of the input object s. We present a theoretical result to justify our diffusion formulation. The sec ond step jointly optimizes the best pose for each object. The final step refines the output of the second step. Experimental results on benchmark datasets of st ructurefrom-motion and scan-based geometry reconstruction show that our approach offers more accurate absolute poses than state-of-the-art pose synchronization techniques.
****
Unsupervised Continual Semantic Adaptation Through Neural Rendering
Zhizheng Liu, Francesco Milano, Jonas Frey, Roland Siegwart, Hermann Blum, Cesar Cadena; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern R ecognition (CVPR), 2023, pp. 3031-3040
An increasing amount of applications rely on data-driven models that are deploye d for perception tasks across a sequence of scenes. Due to the mismatch between training and deployment data, adapting the model on the new scenes is often cruc ial to obtain good performance. In this work, we study continual multi-scene ada ptation for the task of semantic segmentation, assuming that no ground-truth lab els are available during deployment and that performance on the previous scenes should be maintained. We propose training a Semantic-NeRF network for each scene by fusing the predictions of a segmentation model and then using the view-consi stent rendered semantic labels as pseudo-labels to adapt the model. Through join

t training with the segmentation model, the Semantic-NeRF model effectively enab les 2D-3D knowledge transfer. Furthermore, due to its compact size, it can be st ored in a long-term memory and subsequently used to render data from arbitrary v iewpoints to reduce forgetting. We evaluate our approach on ScanNet, where we ou tperform both a voxel-based baseline and a state-of-the-art unsupervised domain adaptation method.
****
Controllable Light Diffusion for Portraits
David Futschik, Kelvin Ritland, James Vecore, Sean Fanello, Sergio Orts-Escolano , Brian Curless, Daniel Sýkora, Rohit Pandey; Proceedings of the IEEE/CVF Confer ence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8412-8421
We introduce light diffusion, a novel method to improve lighting in portraits, s oftening harsh shadows and specular highlights while preserving overall scene il lumination. Inspired by professional photographers' diffusers and scrims, our me thod softens lighting given only a single portrait photo. Previous portrait reli ghting approaches focus on changing the entire lighting environment, removing sh adows (ignoring strong specular highlights), or removing shading entirely. In co ntrast, we propose a learning based method that allows us to control the amount of light diffusion and apply it on in-the-wild portraits. Additionally, we desig n a method to synthetically generate plausible external shadows with sub-surface scattering effects while conforming to the shape of the subject's face. Finally , we show how our approach can increase the robustness of higher level vision ap plications, such as albedo estimation, geometry estimation and semantic segmenta tion.
****
Token Boosting for Robust Self-Supervised Visual Transformer Pre-Training
Tianjiao Li, Lin Geng Foo, Ping Hu, Xindi Shang, Hossein Rahmani, Zehuan Yuan, J un Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Re cognition (CVPR), 2023, pp. 24027-24038
Learning with large-scale unlabeled data has become a powerful tool for pre-trai ning Visual Transformers (VTs). However, prior works tend to overlook that, in r eal-world scenarios, the input data may be corrupted and unreliable. Pre-trainin g VTs on such corrupted data can be challenging, especially when we pre-train vi a the masked autoencoding approach, where both the inputs and masked "ground tru th" targets can potentially be unreliable in this case. To address this limitati on, we introduce the Token Boosting Module (TBM) as a plug-and-play component fo r VTs that effectively allows the VT to learn to extract clean and robust featur es during masked autoencoding pre-training. We provide theoretical analysis to s how how TBM improves model pre-training with more robust and generalizable repre sentations, thus benefiting downstream tasks. We conduct extensive experiments t o analyze TBM's effectiveness, and results on four corrupted datasets demonstrat e that TBM consistently improves performance on downstream tasks.
****
Multi-View Adversarial Discriminator: Mine the Non-Causal Factors for Object Det ection in Unseen Domains
Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, Lei Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp . 8103-8112
Domain shift degrades the performance of object detection models in practical ap plications. To alleviate the influence of domain shift, plenty of previous work try to decouple and learn the domain-invariant (common) features from source dom ains via domain adversarial learning (DAL). However, inspired by causal mechanis ms, we find that previous methods ignore the implicit insignificant non-causal f actors hidden in the common features. This is mainly due to the single-view natu re of DAL. In this work, we present an idea to remove non-causal factors from co mmon features by multi-view adversarial training on source domains, because we o bserve that such insignificant non-causal factors may still be significant in ot her latent spaces (views) due to the multi-mode structure of data. To summarize, we propose a Multi-view Adversarial Discriminator (MAD) based domain generaliza tion model, consisting of a Spurious Correlations Generator (SCG) that increases

 the diversity of source domain by random augmentation and a Multi-View Domain C
lassifier (MVDC) that maps features to multiple latent spaces, such that the non
-causal factors are removed and the domain-invariant features are purified. Exte
nsive experiments on six benchmarks show our MAD obtains state-of-the-art perfor
mance.
****
MaskCon: Masked Contrastive Learning for Coarse-Labelled Dataset
Chen Feng, Ioannis Patras; Proceedings of the IEEE/CVF Conference on Computer Vi
sion and Pattern Recognition (CVPR), 2023, pp. 19913-19922
Deep learning has achieved great success in recent years with the aid of advance
d neural network structures and large-scale human-annotated datasets. However, i
t is often costly and difficult to accurately and efficiently annotate large-sca
le datasets, especially for some specialized domains where fine-grained labels a
re required. In this setting, coarse labels are much easier to acquire as they d
o not require expert knowledge. In this work, we propose a contrastive learning
method, called masked contrastive learning (MaskCon) to address the under-explor
ed problem setting, where we learn with a coarse-labelled dataset in order to ad
dress a finer labelling problem. More specifically, within the contrastive learn
ing framework, for each sample our method generates soft-labels with the aid of
coarse labels against other samples and another augmented view of the sample in
question. By contrast to self-supervised contrastive learning where only the sam
ple's augmentations are considered hard positives, and in supervised contrastive
 learning where only samples with the same coarse labels are considered hard pos
itives, we propose soft labels based on sample distances, that are masked by the
 coarse labels. This allows us to utilize both inter-sample relations and coarse
 labels. We demonstrate that our method can obtain as special cases many existin
g state-of-the-art works and that it provides tighter bounds on the generalizati
on error. Experimentally, our method achieves significant improvement over the c
urrent state-of-the-art in various datasets, including CIFAR10, CIFAR100, ImageN
et-1K, Standford Online Products and Stanford Cars196 datasets. Code and annotat
ions are available at https://github.com/MrChenFeng/MaskCon_CVPR2023.
****
Boosting Low-Data Instance Segmentation by Unsupervised Pre-Training With Salien
cy Prompt
Hao Li, Dingwen Zhang, Nian Liu, Lechao Cheng, Yalun Dai, Chao Zhang, Xinggang W
ang, Junwei Han; Proceedings of the IEEE/CVF Conference on Computer Vision and P
attern Recognition (CVPR), 2023, pp. 15485-15494
Recently, inspired by DETR variants, query-based end-to-end instance segmentatio
n (QEIS) methods have outperformed CNN-based models on large-scale datasets. Yet
 they would lose efficacy when only a small amount of training data is available
 since it's hard for the crucial queries/kernels to learn localization and shape
 priors. To this end, this work offers a novel unsupervised pre-training solutio
n for low-data regimes. Inspired by the recent success of the Prompting techniqu
e, we introduce a new pre-training method that boosts QEIS models by giving Sali
ency Prompt for queries/kernels. Our method contains three parts: 1) Saliency Ma
sks Proposal is responsible for generating pseudo masks from unlabeled images ba
sed on the saliency mechanism. 2) Prompt-Kernel Matching transfers pseudo masks
into prompts and injects the corresponding localization and shape priors to the
best-matched kernels. 3) Kernel Supervision is applied to supply supervision at
the kernel level for robust learning. From a practical perspective, our pre-trai
ning method helps QEIS models achieve a similar convergence speed and comparable
 performance with CNN-based models in low-data regimes. Experimental results sho
w that our method significantly boosts several QEIS models on three datasets.
****
Virtual Occlusions Through Implicit Depth
Jamie Watson, Mohamed Sayed, Zawar Qureshi, Gabriel J. Brostow, Sara Vicente, Oi
sin Mac Aodha, Michael Firman; Proceedings of the IEEE/CVF Conference on Compute
r Vision and Pattern Recognition (CVPR), 2023, pp. 9053-9064
For augmented reality (AR), it is important that virtual assets appear to 'sit a
mong' real world objects. The virtual element should variously occlude and be oc

cluded by real matter, based on a plausible depth ordering. This occlusion shoul d be consistent over time as the viewer's camera moves. Unfortunately, small mis takes in the estimated scene depth can ruin the downstream occlusion mask, and t hereby the AR illusion. Especially in real-time settings, depths inferred near b oundaries or across time can be inconsistent. In this paper, we challenge the ne ed for depth-regression as an intermediate step. We instead propose an implicit model for depth and use that to predict the occlusion mask directly. The inputs to our network are one or more color images, plus the known depths of any virtua l geometry. We show how our occlusion predictions are more accurate and more tem porally stable than predictions derived from traditional depth-estimation models . We obtain state-of-the-art occlusion results on the challenging ScanNetv2 data set and superior qualitative results on real scenes.
****
AGAIN: Adversarial Training With Attribution Span Enlargement and Hybrid Feature Fusion
Shenglin Yin, Kelu Yao, Sheng Shi, Yangzhou Du, Zhen Xiao; Proceedings of the IE EE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2 0544-20553
The deep neural networks (DNNs) trained by adversarial training (AT) usually suf fered from significant robust generalization gap, i.e., DNNs achieve high traini ng robustness but low test robustness. In this paper, we propose a generic metho d to boost the robust generalization of AT methods from the novel perspective of attribution span. To this end, compared with standard DNNs, we discover that th e generalization gap of adversarially trained DNNs is caused by the smaller attr ibution span on the input image. In other words, adversarially trained DNNs tend to focus on specific visual concepts on training images, causing its limitation on test robustness. In this way, to enhance the robustness, we propose an effec tive method to enlarge the learned attribution span. Besides, we use hybrid feat ure statistics for feature fusion to enrich the diversity of features. Extensive experiments show that our method can effectively improves robustness of adversa rially trained DNNs, outperforming previous SOTA methods. Furthermore, we provid e a theoretical analysis of our method to prove its effectiveness.
****
Instance Relation Graph Guided Source-Free Domain Adaptive Object Detection
Vibashan VS, Poojan Oza, Vishal M. Patel; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3520-3530
Unsupervised Domain Adaptation (UDA) is an effective approach to tackle the issu e of domain shift. Specifically, UDA methods try to align the source and target representations to improve generalization on the target domain. Further, UDA met hods work under the assumption that the source data is accessible during the ada ptation process. However, in real-world scenarios, the labelled source data is o ften restricted due to privacy regulations, data transmission constraints, or pr oprietary data concerns. The Source-Free Domain Adaptation (SFDA) setting aims t o alleviate these concerns by adapting a source-trained model for the target dom ain without requiring access to the source data. In this paper, we explore the S FDA setting for the task of adaptive object detection. To this end, we propose a novel training strategy for adapting a source-trained object detector to the ta rget domain without source data. More precisely, we design a novel contrastive l oss to enhance the target representations by exploiting the objects relations fo r a given target domain input. These object instance relations are modelled usin g an Instance Relation Graph (IRG) network, which are then used to guide the con trastive representation learning. In addition, we utilize a student-teacher to e ffectively distill knowledge from source-trained model to target domain. Extensi ve experiments on multiple object detection benchmark datasets show that the pro posed approach is able to efficiently adapt source-trained object detectors to t he target domain, outperforming state-of-the-art domain adaptive detection metho ds. Code and models are provided in https://viudomain.github.io/irg-sfda-web/
****
Instant Multi-View Head Capture Through Learnable Registration
Timo Bolkart, Tianye Li, Michael J. Black; Proceedings of the IEEE/CVF Conferenc

e on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 768-779
Existing methods for capturing datasets of 3D heads in dense semantic correspondence are slow and commonly address the problem in two separate steps; multi-view stereo (MVS) reconstruction followed by non-rigid registration. To simplify this process, we introduce TEMPEH (Towards Estimation of 3D Meshes from Performances of Expressive Heads) to directly infer 3D heads in dense correspondence from calibrated multi-view images. Registering datasets of 3D scans typically requires manual parameter tuning to find the right balance between accurately fitting the scans' surfaces and being robust to scanning noise and outliers. Instead, we propose to jointly register a 3D head dataset while training TEMPEH. Specifically, during training, we minimize a geometric loss commonly used for surface registration, effectively leveraging TEMPEH as a regularizer. Our multi-view head inference builds on a volumetric feature representation that samples and fuses features from each view using camera calibration information. To account for partial occlusions and a large capture volume that enables head movements, we use view- and surface-aware feature fusion, and a spatial transformer-based head localization module, respectively. We use raw MVS scans as supervision during training, but, once trained, TEMPEH directly predicts 3D heads in dense correspondence without requiring scans. Predicting one head takes about 0.3 seconds with a median reconstruction error of 0.26 mm, 64% lower than the current state-of-the-art. This enables the efficient capture of large datasets containing multiple people and diverse facial motions. Code, model, and data are publicly available at https://tempeh.is.tue.mpg.de.
****
DiGA: Distil To Generalize and Then Adapt for Domain Adaptive Semantic Segmentation
Fengyi Shen, Akhil Gurram, Ziyuan Liu, He Wang, Alois Knoll; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15866-15877
Domain adaptive semantic segmentation methods commonly utilize stage-wise training, consisting of a warm-up and a self-training stage. However, this popular approach still faces several challenges in each stage: for warm-up, the widely adopted adversarial training often results in limited performance gain, due to blind feature alignment; for self-training, finding proper categorical thresholds is very tricky. To alleviate these issues, we first propose to replace the adversarial training in the warm-up stage by a novel symmetric knowledge distillation module that only accesses the source domain data and makes the model domain generalizable. Surprisingly, this domain generalizable warm-up model brings substantial performance improvement, which can be further amplified via our proposed cross-domain mixture data augmentation technique. Then, for the self-training stage, we propose a threshold-free dynamic pseudo-label selection mechanism to ease the aforementioned threshold problem and make the model better adapted to the target domain. Extensive experiments demonstrate that our framework achieves remarkable and consistent improvements compared to the prior arts on popular benchmarks. Codes and models are available at https://github.com/fy-vision/DiGA
****
DiffSwap: High-Fidelity and Controllable Face Swapping via 3D-Aware Masked Diffusion
Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8568-8577
In this paper, we propose DiffSwap, a diffusion model based framework for high-fidelity and controllable face swapping. Unlike previous work that relies on carefully designed network architectures and loss functions to fuse the information from the source and target faces, we reformulate the face swapping as a conditional inpainting task, performed by a powerful diffusion model guided by the desired face attributes (e.g., identity and landmarks). An important issue that makes it nontrivial to apply diffusion models to face swapping is that we cannot perform the time-consuming multi-step sampling to obtain the generated image during training. To overcome this, we propose a midpoint estimation method to efficient

ly recover a reasonable diffusion result of the swapped face with only 2 steps, which enables us to introduce identity constraints to improve the face swapping quality. Our framework enjoys several favorable properties more appealing than prior arts: 1) Controllable. Our method is based on conditional masked diffusion on the latent space, where the mask and the conditions can be fully controlled and customized. 2) High-fidelity. The formulation of conditional inpainting can fully exploit the generative ability of diffusion models and can preserve the background of target images with minimal artifacts. 3) Shape-preserving. The controllability of our method enables us to use 3D-aware landmarks as the condition during generation to preserve the shape of the source face. Extensive experiments on both FF++ and FFHQ demonstrate that our method can achieve state-of-the-art face swapping results both qualitatively and quantitatively.
****
GINA-3D: Learning To Generate Implicit Neural Assets in the Wild
Bokui Shen, Xinchen Yan, Charles R. Qi, Mahyar Najibi, Boyang Deng, Leonidas Guibas, Yin Zhou, Dragomir Anguelov; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4913-4926
Modeling the 3D world from sensor data for simulation is a scalable way of developing testing and validation environments for robotic learning problems such as autonomous driving. However, manually creating or re-creating real-world-like environments is difficult, expensive, and not scalable. Recent generative model techniques have shown promising progress to address such challenges by learning 3D assets using only plentiful 2D images -- but still suffer limitations as they leverage either human-curated image datasets or renderings from manually-created synthetic 3D environments. In this paper, we introduce GINA-3D, a generative model that uses real-world driving data from camera and LiDAR sensors to create photo-realistic 3D implicit neural assets of diverse vehicles and pedestrians. Compared to the existing image datasets, the real-world driving setting poses new challenges due to occlusions, lighting-variations and long-tail distributions. GINA-3D tackles these challenges by decoupling representation learning and generative modeling into two stages with a learned tri-plane latent structure, inspired by recent advances in generative modeling of images. To evaluate our approach, we construct a large-scale object-centric dataset containing over 520K images of vehicles and pedestrians from the Waymo Open Dataset, and a new set of 80K images of long-tail instances such as construction equipment, garbage trucks, and cable cars. We compare our model with existing approaches and demonstrate that it achieves state-of-the-art performance in quality and diversity for both generated images and geometries.
****
Consistent Direct Time-of-Flight Video Depth Super-Resolution
Zhanghao Sun, Wei Ye, Jinhui Xiong, Gyeongmin Choe, Jialiang Wang, Shuochen Su, Rakesh Ranjan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5075-5085
Direct time-of-flight (dToF) sensors are promising for next-generation on-device 3D sensing. However, limited by manufacturing capabilities in a compact module, the dToF data has low spatial resolution (e.g., 20x30 for iPhone dToF), and it requires a super-resolution step before being passed to downstream tasks. In this paper, we solve this super-resolution problem by fusing the low-resolution dToF data with the corresponding high-resolution RGB guidance. Unlike the conventional RGB-guided depth enhancement approaches which perform the fusion in a per-frame manner, we propose the first multi-frame fusion scheme to mitigate the spatial ambiguity resulting from the low-resolution dToF imaging. In addition, dToF sensors provide unique depth histogram information for each local patch, and we incorporate this dToF-specific feature in our network design to further alleviate spatial ambiguity. To evaluate our models on complex dynamic indoor environments and to provide a large-scale dToF sensor dataset, we introduce DyDToF, the first synthetic RGB-dToF video dataset that features dynamic objects and a realistic dToF simulator following the physical imaging process. We believe the methods and dataset are beneficial to a broad community as dToF depth sensing is becoming mainstream on mobile devices. Our code and data are publicly available. https

://github.com/facebookresearch/DVSR/
****

Crossing the Gap: Domain Generalization for Image Captioning
Yuchen Ren, Zhendong Mao, Shancheng Fang, Yan Lu, Tong He, Hao Du, Yongdong Zhang, Wanli Ouyang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2871-2880
Existing image captioning methods are under the assumption that the training and testing data are from the same domain or that the data from the target domain (i.e., the domain that testing data lie in) are accessible. However, this assumption is invalid in real-world applications where the data from the target domain is inaccessible. In this paper, we introduce a new setting called Domain Generalization for Image Captioning (DGIC), where the data from the target domain is unseen in the learning process. We first construct a benchmark dataset for DGIC, which helps us to investigate models' domain generalization (DG) ability on unseen domains. With the support of the new benchmark, we further propose a new framework called language-guided semantic metric learning (LSML) for the DGIC setting. Experiments on multiple datasets demonstrate the challenge of the task and the effectiveness of our newly proposed benchmark and LSML framework.
****

Probabilistic Prompt Learning for Dense Prediction
Hyeongjun Kwon, Taeyong Song, Somi Jeong, Jin Kim, Jinhyun Jang, Kwanghoon Sohn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6768-6777
Recent progress in deterministic prompt learning has become a promising alternative to various downstream vision tasks, enabling models to learn powerful visual representations with the help of pre-trained vision-language models. However, this approach results in limited performance for dense prediction tasks that require handling more complex and diverse objects, since a single and deterministic description cannot sufficiently represent the entire image. In this paper, we present a novel probabilistic prompt learning to fully exploit the vision-language knowledge in dense prediction tasks. First, we introduce learnable class-agnostic attribute prompts to describe universal attributes across the object class. The attributes are combined with class information and visual-context knowledge to define the class-specific textual distribution. Text representations are sampled and used to guide the dense prediction task using the probabilistic pixel-text matching loss, enhancing the stability and generalization capability of the proposed method. Extensive experiments on different dense prediction tasks and ablation studies demonstrate the effectiveness of our proposed method.
****

Learned Image Compression With Mixed Transformer-CNN Architectures
Jinming Liu, Heming Sun, Jiro Katto; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14388-14397
Learned image compression (LIC) methods have exhibited promising progress and superior rate-distortion performance compared with classical image compression standards. Most existing LIC methods are Convolutional Neural Networks-based (CNN-based) or Transformer-based, which have different advantages. Exploiting both advantages is a point worth exploring, which has two challenges: 1) how to effectively fuse the two methods? 2) how to achieve higher performance with a suitable complexity? In this paper, we propose an efficient parallel Transformer-CNN Mixture (TCM) block with a controllable complexity to incorporate the local modeling ability of CNN and the non-local modeling ability of transformers to improve the overall architecture of image compression models. Besides, inspired by the recent progress of entropy estimation models and attention modules, we propose a channel-wise entropy model with parameter-efficient swin-transformer-based attention (SWAtten) modules by using channel squeezing. Experimental results demonstrate our proposed method achieves state-of-the-art rate-distortion performances on three different resolution datasets (i.e., Kodak, Tecnick, CLIC Professional Validation) compared to existing LIC methods. The code is at https://github.com/jmliu206/LIC_TCM.
****

Exploring Intra-Class Variation Factors With Learnable Cluster Prompts for Semi-Supervised Image Synthesis
Yunfei Zhang, Xiaoyang Huo, Tianyi Chen, Si Wu, Hau San Wong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7392-7401
Semi-supervised class-conditional image synthesis is typically performed by inferring and injecting class labels into a conditional Generative Adversarial Network (GAN). The supervision in the form of class identity may be inadequate to model classes with diverse visual appearances. In this paper, we propose a Learnable Cluster Prompt-based GAN (LCP-GAN) to capture class-wise characteristics and intra-class variation factors with a broader source of supervision. To exploit partially labeled data, we perform soft partitioning on each class, and explore the possibility of associating intra-class clusters with learnable visual concepts in the feature space of a pre-trained language-vision model, e.g., CLIP. For class-conditional image generation, we design a cluster-conditional generator by injecting a combination of intra-class cluster label embeddings, and further incorporate a real-fake classification head on top of CLIP to distinguish real instances from the synthesized ones, conditioned on the learnable cluster prompts. This significantly strengthens the generator with more semantic language supervision. LCP-GAN not only possesses superior generation capability but also matches the performance of the fully supervised version of the base models: BigGAN and StyleGAN2-ADA, on multiple standard benchmarks.
****
NeAT: Learning Neural Implicit Surfaces With Arbitrary Topologies From Multi-View Images
Xiaoxu Meng, Weikai Chen, Bo Yang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 248-258
Recent progress in neural implicit functions has set new state-of-the-art in reconstructing high-fidelity 3D shapes from a collection of images. However, these approaches are limited to closed surfaces as they require the surface to be represented by a signed distance field. In this paper, we propose NeAT, a new neural rendering framework that can learn implicit surfaces with arbitrary topologies from multi-view images. In particular, NeAT represents the 3D surface as a level set of a signed distance function (SDF) with a validity branch for estimating the surface existence probability at the query positions. We also develop a novel neural volume rendering method, which uses SDF and validity to calculate the volume opacity and avoids rendering points with low validity. NeAT supports easy field-to-mesh conversion using the classic Marching Cubes algorithm. Extensive experiments on DTU, MGN, and Deep Fashion 3D datasets indicate that our approach is able to faithfully reconstruct both watertight and non-watertight surfaces. In particular, NeAT significantly outperforms the state-of-the-art methods in the task of open surface reconstruction both quantitatively and qualitatively.
****
Quantum Multi-Model Fitting
Matteo Farina, Luca Magri, Willi Menapace, Elisa Ricci, Vladislav Golyanik, Federica Arrigoni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13640-13649
Geometric model fitting is a challenging but fundamental computer vision problem. Recently, quantum optimization has been shown to enhance robust fitting for the case of a single model, while leaving the question of multi-model fitting open. In response to this challenge, this paper shows that the latter case can significantly benefit from quantum hardware and proposes the first quantum approach to multi-model fitting (MMF). We formulate MMF as a problem that can be efficiently sampled by modern adiabatic quantum computers without the relaxation of the objective function. We also propose an iterative and decomposed version of our method, which supports real-world-sized problems. The experimental evaluation demonstrates promising results on a variety of datasets. The source code is available at https://github.com/FarinaMatteo/qmmf.
****
SPARF: Neural Radiance Fields From Sparse and Noisy Poses

Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, Federico Tombari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4190-4200
Neural Radiance Field (NeRF) has recently emerged as a powerful representation to synthesize photorealistic novel views. While showing impressive performance, it relies on the availability of dense input views with highly accurate camera poses, thus limiting its application in real-world scenarios. In this work, we introduce Sparse Pose Adjusting Radiance Field (SPARF), to address the challenge of novel-view synthesis given only few wide-baseline input images (as low as 3) with noisy camera poses. Our approach exploits multi-view geometry constraints in order to jointly learn the NeRF and refine the camera poses. By relying on pixel matches extracted between the input views, our multi-view correspondence objective enforces the optimized scene and camera poses to converge to a global and geometrically accurate solution. Our depth consistency loss further encourages the reconstructed scene to be consistent from any viewpoint. Our approach sets a new state of the art in the sparse-view regime on multiple challenging datasets.
****
ABLE-NeRF: Attention-Based Rendering With Learnable Embeddings for Neural Radiance Field
Zhe Jun Tang, Tat-Jen Cham, Haiyu Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16559-16568
Neural Radiance Field (NeRF) is a popular method in representing 3D scenes by optimising a continuous volumetric scene function. Its large success which lies in applying volumetric rendering (VR) is also its Achilles' heel in producing view-dependent effects. As a consequence, glossy and transparent surfaces often appear murky. A remedy to reduce these artefacts is to constrain this VR equation by excluding volumes with back-facing normal. While this approach has some success in rendering glossy surfaces, translucent objects are still poorly represented. In this paper, we present an alternative to the physics-based VR approach by introducing a self-attention-based framework on volumes along a ray. In addition, inspired by modern game engines which utilise Light Probes to store local lighting passing through the scene, we incorporate Learnable Embeddings to capture view dependent effects within the scene. Our method, which we call ABLE-NeRF, significantly reduces 'blurry' glossy surfaces in rendering and produces realistic translucent surfaces which lack in prior art. In the Blender dataset, ABLE-NeRF achieves SOTA results and surpasses Ref-NeRF in all 3 image quality metrics PSNR, SSIM, LPIPS.
****
Local Implicit Normalizing Flow for Arbitrary-Scale Image Super-Resolution
Jie-En Yao, Li-Yuan Tsao, Yi-Chen Lo, Roy Tseng, Chia-Che Chang, Chun-Yi Lee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1776-1785
Flow-based methods have demonstrated promising results in addressing the ill-posed nature of super-resolution (SR) by learning the distribution of high-resolution (HR) images with the normalizing flow. However, these methods can only perform a predefined fixed-scale SR, limiting their potential in real-world applications. Meanwhile, arbitrary-scale SR has gained more attention and achieved great progress. Nonetheless, previous arbitrary-scale SR methods ignore the ill-posed problem and train the model with per-pixel L1 loss, leading to blurry SR outputs. In this work, we propose "Local Implicit Normalizing Flow" (LINF) as a unified solution to the above problems. LINF models the distribution of texture details under different scaling factors with normalizing flow. Thus, LINF can generate photo-realistic HR images with rich texture details in arbitrary scale factors. We evaluate LINF with extensive experiments and show that LINF achieves the state-of-the-art perceptual quality compared with prior arbitrary-scale SR methods.
****
WinCLIP: Zero-/Few-Shot Anomaly Classification and Segmentation
Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, Onkar Dabeer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19606-19616

Visual anomaly classification and segmentation are vital for automating industrial quality inspection. The focus of prior research in the field has been on training custom models for each quality inspection task, which requires task-specific images and annotation. In this paper we move away from this regime, addressing zero-shot and few-normal-shot anomaly classification and segmentation. Recently CLIP, a vision-language model, has shown revolutionary generality with competitive zero/few-shot performance in comparison to full-supervision. But CLIP falls short on anomaly classification and segmentation tasks. Hence, we propose window-based CLIP (WinCLIP) with (1) a compositional ensemble on state words and prompt templates and (2) efficient extraction and aggregation of window/patch/image-level features aligned with text. We also propose its few-normal-shot extension WinCLIP+, which uses complementary information from normal images. In MVTec-AD (and VisA), without further tuning, WinCLIP achieves 91.8%/85.1% (78.1%/79.6%) AUROC in zero-shot anomaly classification and segmentation while WinCLIP+ does 93.1%/95.2% (83.8%/96.4%) in 1-normal-shot, surpassing state-of-the-art by large margins.
****

PermutoSDF: Fast Multi-View Reconstruction With Implicit Surfaces Using Permutohedral Lattices

Radu Alexandru Rosu, Sven Behnke; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8466-8475

Neural radiance-density field methods have become increasingly popular for the task of novel-view rendering. Their recent extension to hash-based positional encoding ensures fast training and inference with visually pleasing results. However, density-based methods struggle with recovering accurate surface geometry. Hybrid methods alleviate this issue by optimizing the density based on an underlying SDF. However, current SDF methods are overly smooth and miss fine geometric details. In this work, we combine the strengths of these two lines of work in a novel hash-based implicit surface representation. We propose improvements to the two areas by replacing the voxel hash encoding with a permutohedral lattice which optimizes faster, especially for higher dimensions. We additionally propose a regularization scheme which is crucial for recovering high-frequency geometric detail. We evaluate our method on multiple datasets and show that we can recover geometric detail at the level of pores and wrinkles while using only RGB images for supervision. Furthermore, using sphere tracing we can render novel views at 30 fps on an RTX 3090. Code is publicly available at https://radualexandru.github.io/permuto_sdf
****

TriDet: Temporal Action Detection With Relative Boundary Modeling

Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18857-18866

In this paper, we present a one-stage framework TriDet for temporal action detection. Existing methods often suffer from imprecise boundary predictions due to the ambiguous action boundaries in videos. To alleviate this problem, we propose a novel Trident-head to model the action boundary via an estimated relative probability distribution around the boundary. In the feature pyramid of TriDet, we propose a Scalable-Granularity Perception (SGP) layer to aggregate information across different temporal granularities, which is much more efficient than the recent transformer-based feature pyramid. Benefiting from the Trident-head and the SGP-based feature pyramid, TriDet achieves state-of-the-art performance on three challenging benchmarks: THUMOS14, HACS and EPIC-KITCHEN 100, with lower computational costs, compared to previous methods. For example, TriDet hits an average mAP of 69.3% on THUMOS14, outperforming the previous best by 2.5%, but with only 74.6% of its latency.
****

Detection Hub: Unifying Object Detection Datasets via Query Adaptation on Language Embedding

Lingchen Meng, Xiyang Dai, Yinpeng Chen, Pengchuan Zhang, Dongdong Chen, Mengchen Liu, Jianfeng Wang, Zuxuan Wu, Lu Yuan, Yu-Gang Jiang; Proceedings of the IEEE

/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11402-11411

Combining multiple datasets enables performance boost on many computer vision tasks. But similar trend has not been witnessed in object detection when combining multiple datasets due to two inconsistencies among detection datasets: taxonomy difference and domain gap. In this paper, we address these challenges by a new design (named Detection Hub) that is dataset-aware and category-aligned. It not only mitigates the dataset inconsistency but also provides coherent guidance for the detector to learn across multiple datasets. In particular, the dataset-aware design is achieved by learning a dataset embedding that is used to adapt object queries as well as convolutional kernels in detection heads. The categories across datasets are semantically aligned into a unified space by replacing one-hot category representations with word embedding and leveraging the semantic coherence of language embedding. Detection Hub fulfills the benefits of large data on object detection. Experiments demonstrate that joint training on multiple datasets achieves significant performance gains over training on each dataset alone. Detection Hub further achieves SoTA performance on UODB benchmark with wide variety of datasets.

****

Dream3D: Zero-Shot Text-to-3D Synthesis Using 3D Shape Prior and Text-to-Image Diffusion Models

Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, Shenghua Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20908-20918

Recent CLIP-guided 3D optimization methods, such as DreamFields and PureCLIPNeRF, have achieved impressive results in zero-shot text-to-3D synthesis. However, due to scratch training and random initialization without prior knowledge, these methods often fail to generate accurate and faithful 3D structures that conform to the input text. In this paper, we make the first attempt to introduce explicit 3D shape priors into the CLIP-guided 3D optimization process. Specifically, we first generate a high-quality 3D shape from the input text in the text-to-shape stage as a 3D shape prior. We then use it as the initialization of a neural radiance field and optimize it with the full prompt. To address the challenging text-to-shape generation task, we present a simple yet effective approach that directly bridges the text and image modalities with a powerful text-to-image diffusion model. To narrow the style domain gap between the images synthesized by the text-to-image diffusion model and shape renderings used to train the image-to-shape generator, we further propose to jointly optimize a learnable text prompt and fine-tune the text-to-image diffusion model for rendering-style image generation. Our method, Dream3D, is capable of generating imaginative 3D content with superior visual quality and shape accuracy compared to state-of-the-art methods. Our project page is at https://bluestyle97.github.io/dream3d/.

****

Adversarial Normalization: I Can Visualize Everything (ICE)

Hoyoung Choi, Seungwan Jin, Kyungsik Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12115-12124

Vision transformers use [CLS] tokens to predict image classes. Their explainability visualization has been studied using relevant information from [CLS] tokens or focusing on attention scores during self-attention. Such visualization, however, is challenging because of the dependence of the structure of a vision transformer on skip connections and attention operators, the instability of non-linearities in the learning process, and the limited reflection of self-attention scores on relevance. We argue that the output vectors for each input patch token in a vision transformer retain the image information of each patch location, which can facilitate the prediction of an image class. In this paper, we propose ICE (Adversarial Normalization: I Can visualize Everything), a novel method that enables a model to directly predict a class for each patch in an image; thus, advancing the effective visualization of the explainability of a vision transformer. Our method distinguishes background from foreground regions by predicting background classes for patches that do not determine image classes. We used the DeiT-S

model, the most representative model employed in studies, on the explainability visualization of vision transformers. On the ImageNet-Segmentation dataset, ICE outperformed all explainability visualization methods for four cases depending on the model size. We also conducted quantitative and qualitative analyses on the tasks of weakly-supervised object localization and unsupervised object discovery. On the CUB-200-2011 and PASCALVOC07/12 datasets, ICE achieved comparable performance to the state-of-the-art methods. We incorporated ICE into the encoder of DeiT-S and improved efficiency by 44.01% on the ImageNet dataset over that achieved by the original DeiT-S model. We showed performance on the accuracy and efficiency comparable to EViT, the state-of-the-art pruning model, demonstrating the effectiveness of ICE. The code is available at https://github.com/Hanyang-HCC-Lab/ICE.
****
Reinforcement Learning-Based Black-Box Model Inversion Attacks
Gyojin Han, Jaehyun Choi, Haeil Lee, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20504-20513
Model inversion attacks are a type of privacy attack that reconstructs private data used to train a machine learning model, solely by accessing the model. Recently, white-box model inversion attacks leveraging Generative Adversarial Networks (GANs) to distill knowledge from public datasets have been receiving great attention because of their excellent attack performance. On the other hand, current black-box model inversion attacks that utilize GANs suffer from issues such as being unable to guarantee the completion of the attack process within a predetermined number of query accesses or achieve the same level of performance as white-box attacks. To overcome these limitations, we propose a reinforcement learning-based black-box model inversion attack. We formulate the latent space search as a Markov Decision Process (MDP) problem and solve it with reinforcement learning. Our method utilizes the confidence scores of the generated images to provide rewards to an agent. Finally, the private data can be reconstructed using the latent vectors found by the agent trained in the MDP. The experiment results on various datasets and models demonstrate that our attack successfully recovers the private information of the target model by achieving state-of-the-art attack performance. We emphasize the importance of studies on privacy-preserving machine learning by proposing a more advanced black-box model inversion attack.
****
Learning a Deep Color Difference Metric for Photographic Images
Haoyu Chen, Zhihua Wang, Yang Yang, Qilin Sun, Kede Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22242-22251
Most well-established and widely used color difference (CD) metrics are handcrafted and subject-calibrated against uniformly colored patches, which do not generalize well to photographic images characterized by natural scene complexities. Constructing CD formulae for photographic images is still an active research topic in imaging/illumination, vision science, and color science communities. In this paper, we aim to learn a deep CD metric for photographic images with four desirable properties. First, it well aligns with the observations in vision science that color and form are linked inextricably in visual cortical processing. Second, it is a proper metric in the mathematical sense. Third, it computes accurate CDs between photographic images, differing mainly in color appearances. Fourth, it is robust to mild geometric distortions (e.g., translation or due to parallax), which are often present in photographic images of the same scene captured by different digital cameras. We show that all these properties can be satisfied at once by learning a multi-scale autoregressive normalizing flow for feature transform, followed by the Euclidean distance which is linearly proportional to the human perceptual CD. Quantitative and qualitative experiments on the large-scale SPCD dataset demonstrate the promise of the learned CD metric.
****
1000 FPS HDR Video With a Spike-RGB Hybrid Camera
Yakun Chang, Chu Zhou, Yuchen Hong, Liwen Hu, Chao Xu, Tiejun Huang, Boxin Shi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio

n (CVPR), 2023, pp. 22180-22190
Capturing high frame rate and high dynamic range (HFR&HDR) color videos in high-speed scenes with conventional frame-based cameras is very challenging. The increasing frame rate is usually guaranteed by using shorter exposure time so that the captured video is severely interfered by noise. Alternating exposures could alleviate the noise issue but sacrifice frame rate due to involving long-exposure frames. The neuromorphic spiking camera records high-speed scenes of high dynamic range without colors using a completely different sensing mechanism and visual representation. We introduce a hybrid camera system composed of a spiking and an alternating-exposure RGB camera to capture HFR&HDR scenes with high fidelity. Our insight is to bring each camera's superiority into full play. The spike frames, with accurate fast motion information encoded, are first reconstructed for motion representation, from which the spike-based optical flows guide the recovery of missing temporal information for middle- and long-exposure RGB images while retaining their reliable color appearances. With the strong temporal constraint estimated from spike trains, both missing and distorted colors cross RGB frames are recovered to generate time-consistent and HFR color frames. We collect a new Spike-RGB dataset that contains 300 sequences of synthetic data and 20 groups of real-world data to demonstrate 1000 FPS HDR videos outperforming HDR video reconstruction methods and commercial high-speed cameras.
****
DINN360: Deformable Invertible Neural Network for Latitude-Aware 360deg Image Rescaling
Yichen Guo, Mai Xu, Lai Jiang, Leonid Sigal, Yunjin Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21519-21528
With the rapid development of virtual reality, 360deg images have gained increasing popularity. Their wide field of view necessitates high resolution to ensure image quality. This, however, makes it harder to acquire, store and even process such 360deg images. To alleviate this issue, we propose the first attempt at 360deg image rescaling, which refers to downscaling a 360deg image to a visually valid low-resolution (LR) counterpart and then upscaling to a high-resolution (HR) 360deg image given the LR variant. Specifically, we first analyze two 360deg image datasets and observe several findings that characterize how 360deg images typically change along their latitudes. Inspired by these findings, we propose a novel deformable invertible neural network (INN), named DINN360, for latitude-aware 360deg image rescaling. In DINN360, a deformable INN is designed to downscale the LR image, and project the high-frequency (HF) component to the latent space by adaptively handling various deformations occurring at different latitude regions. Given the downscaled LR image, the high-quality HR image is then reconstructed in a conditional latitude-aware manner by recovering the structure-related HF component from the latent space. Extensive experiments over four public datasets show that our DINN360 method performs considerably better than other state-of-the-art methods for 2x, 4x and 8x 360deg image rescaling.
****
Learning Geometric-Aware Properties in 2D Representation Using Lightweight CAD Models, or Zero Real 3D Pairs
Pattaramanee Arsomngern, Sarana Nutanong, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21371-21381
Cross-modal training using 2D-3D paired datasets, such as those containing multi-view images and 3D scene scans, presents an effective way to enhance 2D scene understanding by introducing geometric and view-invariance priors into 2D features. However, the need for large-scale scene datasets can impede scalability and further improvements. This paper explores an alternative learning method by leveraging a lightweight and publicly available type of 3D data in the form of CAD models. We construct a 3D space with geometric-aware alignment where the similarity in this space reflects the geometric similarity of CAD models based on the Chamfer distance. The acquired geometric-aware properties are then induced into 2D features, which boost performance on downstream tasks more effectively than exis

ting RGB-CAD approaches. Our technique is not limited to paired RGB-CAD datasets. By training exclusively on pseudo pairs generated from CAD-based reconstruction methods, we enhance the performance of SOTA 2D pre-trained models that use ResNet-50 or ViT-B backbones on various 2D understanding tasks. We also achieve comparable results to SOTA methods trained on scene scans on four tasks in NYUv2, SUNRGB-D, indoor ADE20k, and indoor/outdoor COCO, despite using lightweight CAD models or pseudo data.
****
Texts as Images in Prompt Tuning for Multi-Label Image Recognition
Zixian Guo, Bowen Dong, Zhilong Ji, Jinfeng Bai, Yiwen Guo, Wangmeng Zuo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2808-2817
Prompt tuning has been employed as an efficient way to adapt large vision-language pre-trained models (e.g. CLIP) to various downstream tasks in data-limited or label-limited settings. Nonetheless, visual data (e.g., images) is by default prerequisite for learning prompts in existing methods. In this work, we advocate that the effectiveness of image-text contrastive learning in aligning the two modalities (for training CLIP) further makes it feasible to treat texts as images for prompt tuning and introduce TaI prompting. In contrast to the visual data, text descriptions are easy to collect, and their class labels can be directly derived. Particularly, we apply TaI prompting to multi-label image recognition, where sentences in the wild serve as alternatives to images for prompt tuning. Moreover, with TaI, double-grained prompt tuning (TaI-DPT) is further presented to extract both coarse-grained and fine-grained embeddings for enhancing the multi-label recognition performance. Experimental results show that our proposed TaI-DPT outperforms zero-shot CLIP by a large margin on multiple benchmarks, e.g., MS-COCO, VOC2007, and NUS-WIDE, while it can be combined with existing methods of prompting from images to improve recognition performance further. The code is released at https://github.com/guozix/TaI-DPT.
****
Self-Correctable and Adaptable Inference for Generalizable Human Pose Estimation
Zhehan Kan, Shuoshuo Chen, Ce Zhang, Yushun Tang, Zhihai He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5537-5546
A central challenge in human pose estimation, as well as in many other machine learning and prediction tasks, is the generalization problem. The learned network does not have the capability to characterize the prediction error, generate feedback information from the test sample, and correct the prediction error on the fly for each individual test sample, which results in degraded performance in generalization. In this work, we introduce a self-correctable and adaptable inference (SCAI) method to address the generalization challenge of network prediction and use human pose estimation as an example to demonstrate its effectiveness and performance. We learn a correction network to correct the prediction result conditioned by a fitness feedback error. This feedback error is generated by a learned fitness feedback network which maps the prediction result to the original input domain and compares it against the original input. Interestingly, we find that this self-referential feedback error is highly correlated with the actual prediction error. This strong correlation suggests that we can use this error as feedback to guide the correction process. It can be also used as a loss function to quickly adapt and optimize the correction network during the inference process. Our extensive experimental results on human pose estimation demonstrate that the proposed SCAI method is able to significantly improve the generalization capability and performance of human pose estimation.
****
Few-Shot Learning With Visual Distribution Calibration and Cross-Modal Distribution Alignment
Runqi Wang, Hao Zheng, Xiaoyue Duan, Jianzhuang Liu, Yuning Lu, Tian Wang, Songcen Xu, Baochang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23445-23454
Pre-trained vision-language models have inspired much research on few-shot learn

ing. However, with only a few training images, there exist two crucial problems: (1) the visual feature distributions are easily distracted by class-irrelevant information in images, and (2) the alignment between the visual and language feature distributions is difficult. To deal with the distraction problem, we propose a Selective Attack module, which consists of trainable adapters that generate spatial attention maps of images to guide the attacks on class-irrelevant image areas. By messing up these areas, the critical features are captured and the visual distributions of image features are calibrated. To better align the visual and language feature distributions that describe the same object class, we propose a cross-modal distribution alignment module, in which we introduce a vision-language prototype for each class to align the distributions, and adopt the Earth Mover's Distance (EMD) to optimize the prototypes. For efficient computation, the upper bound of EMD is derived. In addition, we propose an augmentation strategy to increase the diversity of the images and the text prompts, which can reduce overfitting to the few-shot training images. Extensive experiments on 11 datasets demonstrate that our method consistently outperforms prior arts in few-shot learning.
****
Referring Multi-Object Tracking
Dongming Wu, Wencheng Han, Tiancai Wang, Xingping Dong, Xiangyu Zhang, Jianbing Shen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14633-14642
Existing referring understanding tasks tend to involve the detection of a single text-referred object. In this paper, we propose a new and general referring understanding task, termed referring multi-object tracking (RMOT). Its core idea is to employ a language expression as a semantic cue to guide the prediction of multi-object tracking. To the best of our knowledge, it is the first work to achieve an arbitrary number of referent object predictions in videos. To push forward RMOT, we construct one benchmark with scalable expressions based on KITTI, named Refer-KITTI. Specifically, it provides 18 videos with 818 expressions, and each expression in a video is annotated with an average of 10.7 objects. Further, we develop a transformer-based architecture TransRMOT to tackle the new task in an online manner, which achieves impressive detection performance and outperforms other counterparts. The Refer-KITTI dataset and the code are released at https://referringmot.github.io.
****
Finetune Like You Pretrain: Improved Finetuning of Zero-Shot Vision Models
Sachin Goyal, Ananya Kumar, Sankalp Garg, Zico Kolter, Aditi Raghunathan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19338-19347
Finetuning image-text models such as CLIP achieves state-of-the-art accuracies on a variety of benchmarks. However, recent works (Kumar et al., 2022; Wortsman et al., 2021) have shown that even subtle differences in the finetuning process can lead to surprisingly large differences in the final performance, both for in-distribution (ID) and out-of-distribution (OOD) data. In this work, we show that a natural and simple approach of mimicking contrastive pretraining consistently outperforms alternative finetuning approaches. Specifically, we cast downstream class labels as text prompts and continue optimizing the contrastive loss between image embeddings and class-descriptive prompt embeddings (contrastive finetuning). Our method consistently outperforms baselines across 7 distribution shift, 6 transfer learning, and 3 few-shot learning benchmarks. On WILDS-iWILDCam, our proposed approach FLYP outperforms the top of the leaderboard by 2.3% ID and 2.7% OOD, giving the highest reported accuracy. Averaged across 7 OOD datasets (2 WILDS and 5 ImageNet associated shifts), FLYP gives gains of 4.2% OOD over standard finetuning and outperforms current state-ofthe-art (LP-FT) by more than 1% both ID and OOD. Similarly, on 3 few-shot learning benchmarks, FLYP gives gains up to 4.6% over standard finetuning and 4.4% over the state-of-the-art. Thus we establish our proposed method of contrastive finetuning as a simple and intuitive state-ofthe-art for supervised finetuning of image-text models like CLIP. Code is available at https://github.com/locuslab/FLYP.

****
GradMA: A Gradient-Memory-Based Accelerated Federated Learning With Alleviated Catastrophic Forgetting
Kangyang Luo, Xiang Li, Yunshi Lan, Ming Gao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3708-3717
Federated Learning (FL) has emerged as a de facto machine learning area and received rapid increasing research interests from the community. However, catastrophic forgetting caused by data heterogeneity and partial participation poses distinctive challenges for FL, which are detrimental to the performance. To tackle the problems, we propose a new FL approach (namely GradMA), which takes inspiration from continual learning to simultaneously correct the server-side and worker-side update directions as well as take full advantage of server's rich computing and memory resources. Furthermore, we elaborate a memory reduction strategy to enable GradMA to accommodate FL with a large scale of workers. We then analyze convergence of GradMA theoretically under the smooth non-convex setting and show that its convergence rate achieves a linear speed up w.r.t the increasing number of sampled active workers. At last, our extensive experiments on various image classification tasks show that GradMA achieves significant performance gains in accuracy and communication efficiency compared to SOTA baselines. We provide our code here: https://github.com/lkyddd/GradMA.
****
Weakly Supervised Temporal Sentence Grounding With Uncertainty-Guided Self-Training
Yifei Huang, Lijin Yang, Yoichi Sato; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18908-18918
The task of weakly supervised temporal sentence grounding aims at finding the corresponding temporal moments of a language description in the video, given video-language correspondence only at video-level. Most existing works select mismatched video-language pairs as negative samples and train the model to generate better positive proposals that are distinct from the negative ones. However, due to the complex temporal structure of videos, proposals distinct from the negative ones may correspond to several video segments but not necessarily the correct ground truth. To alleviate this problem, we propose an uncertainty-guided self-training technique to provide extra self-supervision signals to guide the weakly-supervised learning. The self-training process is based on teacher-student mutual learning with weak-strong augmentation, which enables the teacher network to generate relatively more reliable outputs compared to the student network, so that the student network can learn from the teacher's output. Since directly applying existing self-training methods in this task easily causes error accumulation, we specifically design two techniques in our self-training method: (1) we construct a Bayesian teacher network, leveraging its uncertainty as a weight to suppress the noisy teacher supervisory signals; (2) we leverage the cycle consistency brought by temporal data augmentation to perform mutual learning between the two networks. Experiments demonstrate our method's superiority on Charades-STA and ActivityNet Captions datasets. We also show in the experiment that our self-training method can be applied to improve the performance of multiple backbone methods.
****
Hint-Aug: Drawing Hints From Foundation Vision Transformers Towards Boosted Few-Shot Parameter-Efficient Tuning
Zhongzhi Yu, Shang Wu, Yonggan Fu, Shunyao Zhang, Yingyan (Celine) Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11102-11112
Despite the growing demand for tuning foundation vision transformers (FViTs) on downstream tasks, fully unleashing FViTs' potential under data-limited scenarios (e.g., few-shot tuning) remains a challenge due to FViTs' data-hungry nature. Common data augmentation techniques fall short in this context due to the limited features contained in the few-shot tuning data. To tackle this challenge, we first identify an opportunity for FViTs in few-shot tuning: pretrained FViTs themselves have already learned highly representative features from large-scale pretr

aining data, which are fully preserved during widely used parameter-efficient tu
ning. We thus hypothesize that leveraging those learned features to augment the
tuning data can boost the effectiveness of few-shot FViT tuning. To this end, we
 propose a framework called Hint-based Data Augmentation (Hint-Aug), which aims
to boost FViT in few-shot tuning by augmenting the over-fitted parts of tuning s
amples with the learned features of pretrained FViTs. Specifically, Hint-Aug int
egrates two key enablers: (1) an Attentive Over-fitting Detector (AOD) to detect
 over-confident patches of foundation ViTs for potentially alleviating their ove
r-fitting on the few-shot tuning data and (2) a Confusion-based Feature Infusion
 (CFI) module to infuse easy-to-confuse features from the pretrained FViTs with
the over-confident patches detected by the above AOD in order to enhance the fea
ture diversity during tuning. Extensive experiments and ablation studies on five
 datasets and three parameter-efficient tuning techniques consistently validate
Hint-Aug's effectiveness: 0.04% 32.91% higher accuracy over the state-of-the-art
 (SOTA) data augmentation method under various low-shot settings. For example, o
n the Pet dataset, Hint-Aug achieves a 2.22% higher accuracy with 50% less train
ing data over SOTA data augmentation methods.
****
A Strong Baseline for Generalized Few-Shot Semantic Segmentation
Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, Jose Dolz; Proceedings of the IEE
E/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11
269-11278
This paper introduces a generalized few-shot segmentation framework with a strai
ghtforward training process and an easy-to-optimize inference phase. In particul
ar, we propose a simple yet effective model based on the well-known InfoMax prin
ciple, where the Mutual Information (MI) between the learned feature representat
ions and their corresponding predictions is maximized. In addition, the terms de
rived from our MI-based formulation are coupled with a knowledge distillation te
rm to retain the knowledge on base classes. With a simple training process, our
inference model can be applied on top of any segmentation network trained on bas
e classes. The proposed inference yields substantial improvements on the popular
 few-shot segmentation benchmarks, PASCAL-5^i and COCO-20^i. Particularly, for n
ovel classes, the improvement gains range from 7% to 26% (PASCAL-5^i) and from 3
% to 12% (COCO-20^i) in the 1-shot and 5-shot scenarios, respectively. Furthermo
re, we propose a more challenging setting, where performance gaps are further ex
acerbated. Our code is publicly available at https://github.com/sinahmr/DIaM.
****
AutoRecon: Automated 3D Object Discovery and Reconstruction
Yuang Wang, Xingyi He, Sida Peng, Haotong Lin, Hujun Bao, Xiaowei Zhou; Proceedi
ngs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)
, 2023, pp. 21382-21391
A fully automated object reconstruction pipeline is crucial for digital content
creation. While the area of 3D reconstruction has witnessed profound development
s, the removal of background to obtain a clean object model still relies on diff
erent forms of manual labor, such as bounding box labeling, mask annotations, an
d mesh manipulations. In this paper, we propose a novel framework named AutoReco
n for the automated discovery and reconstruction of an object from multi-view im
ages. We demonstrate that foreground objects can be robustly located and segment
ed from SfM point clouds by leveraging self-supervised 2D vision transformer fea
tures. Then, we reconstruct decomposed neural scene representations with dense s
upervision provided by the decomposed point clouds, resulting in accurate object
 reconstruction and segmentation. Experiments on the DTU, BlendedMVS and CO3D-V2
 datasets demonstrate the effectiveness and robustness of AutoRecon. The code an
d supplementary material are available on the project page: https://zju3dv.githu
b.io/autorecon/.
****
POTTER: Pooling Attention Transformer for Efficient Human Mesh Recovery
Ce Zheng, Xianpeng Liu, Guo-Jun Qi, Chen Chen; Proceedings of the IEEE/CVF Confe
rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1611-1620
Transformer architectures have achieved SOTA performance on the human mesh recov

ery (HMR) from monocular images. However, the performance gain has come at the cost of substantial memory and computational overhead. A lightweight and efficient model to reconstruct accurate human mesh is needed for real-world applications. In this paper, we propose a pure transformer architecture named POoling aTtention TransformER (POTTER) for the HMR task from single images. Observing that the conventional attention module is memory and computationally expensive, we propose an efficient pooling attention module, which significantly reduces the memory and computational cost without sacrificing performance. Furthermore, we design a new transformer architecture by integrating a High-Resolution (HR) stream for the HMR task. The high-resolution local and global features from the HR stream can be utilized for recovering more accurate human mesh. Our POTTER outperforms the SOTA method METRO by only requiring 7% of total parameters and 14% of the Multiply-Accumulate Operations on the Human3.6M (PA-MPJPE) and 3DPW (all three metrics) datasets. Code will be publicly available.
****
Learning a Practical SDR-to-HDRTV Up-Conversion Using New Dataset and Degradation Models
Cheng Guo, Leidong Fan, Ziyu Xue, Xiuhua Jiang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22231-22241
In media industry, the demand of SDR-to-HDRTV up-conversion arises when users possess HDR-WCG (high dynamic range-wide color gamut) TVs while most off-the-shelf footage is still in SDR (standard dynamic range). The research community has started tackling this low-level vision task by learning-based approaches. When applied to real SDR, yet, current methods tend to produce dim and desaturated result, making nearly no improvement on viewing experience. Different from other network-oriented methods, we attribute such deficiency to training set (HDR-SDR pair). Consequently, we propose new HDRTV dataset (dubbed HDRTV4K) and new HDR-to-SDR degradation models. Then, it's used to train a luminance-segmented network (LSN) consisting of a global mapping trunk, and two Transformer branches on bright and dark luminance range. We also update assessment criteria by tailored metrics and subjective experiment. Finally, ablation studies are conducted to prove the effectiveness. Our work is available at: https://github.com/AndreGuo/HDRTVDM.
****
Learning Detailed Radiance Manifolds for High-Fidelity and 3D-Consistent Portrait Synthesis From Monocular Image
Yu Deng, Baoyuan Wang, Heung-Yeung Shum; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4423-4433
A key challenge for novel view synthesis of monocular portrait images is 3D consistency under continuous pose variations. Most existing methods rely on 2D generative models which often leads to obvious 3D inconsistency artifacts. We present a 3D-consistent novel view synthesis approach for monocular portrait images based on a recent proposed 3D-aware GAN, namely Generative Radiance Manifolds (GRAM), which has shown strong 3D consistency at multiview image generation of virtual subjects via the radiance manifolds representation. However, simply learning an encoder to map a real image into the latent space of GRAM can only reconstruct coarse radiance manifolds without faithful fine details, while improving the reconstruction fidelity via instance-specific optimization is time-consuming. We introduce a novel detail manifolds reconstructor to learn 3D-consistent fine details on the radiance manifolds from monocular images, and combine them with the coarse radiance manifolds for high-fidelity reconstruction. The 3D priors derived from the coarse radiance manifolds are used to regulate the learned details to ensure reasonable synthesized results at novel views. Trained on in-the-wild 2D images, our method achieves high-fidelity and 3D-consistent portrait synthesis largely outperforming the prior art. Project page: https://yudeng.github.io/GRAMInverter/
****
Patch-Craft Self-Supervised Training for Correlated Image Denoising
Gregory Vaksman, Michael Elad; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5795-5804
Supervised neural networks are known to achieve excellent results in various ima

ge restoration tasks. However, such training requires datasets composed of pairs of corrupted images and their corresponding ground truth targets. Unfortunately, such data is not available in many applications. For the task of image denoising in which the noise statistics is unknown, several self-supervised training methods have been proposed for overcoming this difficulty. Some of these require knowledge of the noise model, while others assume that the contaminating noise is uncorrelated, both assumptions are too limiting for many practical needs. This work proposes a novel self-supervised training technique suitable for the removal of unknown correlated noise. The proposed approach neither requires knowledge of the noise model nor access to ground truth targets. The input to our algorithm consists of easily captured bursts of noisy shots. Our algorithm constructs artificial patch-craft images from these bursts by patch matching and stitching, and the obtained crafted images are used as targets for the training. Our method does not require registration of the different images within the burst. We evaluate the proposed framework through extensive experiments with synthetic and real image noise.
****
Learning To Fuse Monocular and Multi-View Cues for Multi-Frame Depth Estimation in Dynamic Scenes
Rui Li, Dong Gong, Wei Yin, Hao Chen, Yu Zhu, Kaixuan Wang, Xiaozhi Chen, Jinqiu Sun, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21539-21548
Multi-frame depth estimation generally achieves high accuracy relying on the multi-view geometric consistency. When applied in dynamic scenes, e.g., autonomous driving, this consistency is usually violated in the dynamic areas, leading to corrupted estimations. Many multi-frame methods handle dynamic areas by identifying them with explicit masks and compensating the multi-view cues with monocular cues represented as local monocular depth or features. The improvements are limited due to the uncontrolled quality of the masks and the underutilized benefits of the fusion of the two types of cues. In this paper, we propose a novel method to learn to fuse the multi-view and monocular cues encoded as volumes without needing the heuristically crafted masks. As unveiled in our analyses, the multi-view cues capture more accurate geometric information in static areas, and the monocular cues capture more useful contexts in dynamic areas. To let the geometric perception learned from multi-view cues in static areas propagate to the monocular representation in dynamic areas and let monocular cues enhance the representation of multi-view cost volume, we propose a cross-cue fusion (CCF) module, which includes the cross-cue attention (CCA) to encode the spatially non-local relative intra-relations from each source to enhance the representation of the other. Experiments on real-world datasets prove the significant effectiveness and generalization ability of the proposed method.
****
DynaFed: Tackling Client Data Heterogeneity With Global Dynamics
Renjie Pi, Weizhong Zhang, Yueqi Xie, Jiahui Gao, Xiaoyu Wang, Sunghun Kim, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12177-12186
The Federated Learning (FL) paradigm is known to face challenges under heterogeneous client data. Local training on non-iid distributed data results in deflected local optimum, which causes the client models drift further away from each other and degrades the aggregated global model's performance. A natural solution is to gather all client data onto the server, such that the server has a global view of the entire data distribution. Unfortunately, this reduces to regular training, which compromises clients' privacy and conflicts with the purpose of FL. In this paper, we put forth an idea to collect and leverage global knowledge on the server without hindering data privacy. We unearth such knowledge from the dynamics of the global model's trajectory. Specifically, we first reserve a short trajectory of global model snapshots on the server. Then, we synthesize a small pseudo dataset such that the model trained on it mimics the dynamics of the reserved global model trajectory. Afterward, the synthesized data is used to help aggregate the deflected clients into the global model. We name our method DynaFed, w

hich enjoys the following advantages: 1) we do not rely on any external on-serve
r dataset, which requires no additional cost for data collection; 2) the pseudo
data can be synthesized in early communication rounds, which enables DynaFed to
take effect early for boosting the convergence and stabilizing training; 3) the
pseudo data only needs to be synthesized once and can be directly utilized on th
e server to help aggregation in subsequent rounds. Experiments across extensive
benchmarks are conducted to showcase the effectiveness of DynaFed. We also provi
de insights and understanding of the underlying mechanism of our method.
****
Bias-Eliminating Augmentation Learning for Debiased Federated Learning
Yuan-Yi Xu, Ci-Siang Lin, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20442-20452
Learning models trained on biased datasets tend to observe correlations between
categorical and undesirable features, which result in degraded performances. Mos
t existing debiased learning models are designed for centralized machine learnin
g, which cannot be directly applied to distributed settings like federated learn
ing (FL), which collects data at distinct clients with privacy preserved. To tac
kle the challenging task of debiased federated learning, we present a novel FL f
ramework of Bias-Eliminating Augmentation Learning (FedBEAL), which learns to de
ploy Bias-Eliminating Augmenters (BEA) for producing client-specific bias-confli
cting samples at each client. Since the bias types or attributes are not known i
n advance, a unique learning strategy is presented to jointly train BEA with the
 proposed FL framework. Extensive image classification experiments on datasets w
ith various bias types confirm the effectiveness and applicability of our FedBEA
L, which performs favorably against state-of-the-art debiasing and FL methods fo
r debiased FL.
****
DistilPose: Tokenized Pose Regression With Heatmap Distillation
Suhang Ye, Yingyi Zhang, Jie Hu, Liujuan Cao, Shengchuan Zhang, Lei Shen, Jun Wa
ng, Shouhong Ding, Rongrong Ji; Proceedings of the IEEE/CVF Conference on Comput
er Vision and Pattern Recognition (CVPR), 2023, pp. 2163-2172
In the field of human pose estimation, regression-based methods have been domina
ted in terms of speed, while heatmap-based methods are far ahead in terms of per
formance. How to take advantage of both schemes remains a challenging problem. I
n this paper, we propose a novel human pose estimation framework termed DistilPo
se, which bridges the gaps between heatmap-based and regression-based methods. S
pecifically, DistilPose maximizes the transfer of knowledge from the teacher mod
el (heatmap-based) to the student model (regression-based) through Token-distill
ing Encoder (TDE) and Simulated Heatmaps. TDE aligns the feature spaces of heatm
ap-based and regression-based models by introducing tokenization, while Simulate
d Heatmaps transfer explicit guidance (distribution and confidence) from teacher
 heatmaps into student models. Extensive experiments show that the proposed Dist
ilPose can significantly improve the performance of the regression-based models
while maintaining efficiency. Specifically, on the MSCOCO validation dataset, Di
stilPose-S obtains 71.6% mAP with 5.36M parameter, 2.38 GFLOPs and 40.2 FPS, whi
ch saves 12.95x, 7.16x computational cost and is 4.9x faster than its teacher mo
del with only 0.9 points performance drop. Furthermore, DistilPose-L obtains 74.
4% mAP on MSCOCO validation dataset, achieving a new state-of-the-art among pred
ominant regression-based models.
****
Understanding the Robustness of 3D Object Detection With Bird's-Eye-View Represe
ntations in Autonomous Driving
Zijian Zhu, Yichi Zhang, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Z
hong, Shibao Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision an
d Pattern Recognition (CVPR), 2023, pp. 21600-21610
3D object detection is an essential perception task in autonomous driving to und
erstand the environments. The Bird's-Eye-View (BEV) representations have signifi
cantly improved the performance of 3D detectors with camera inputs on popular be
nchmarks. However, there still lacks a systematic understanding of the robustnes
s of these vision-dependent BEV models, which is closely related to the safety o

f autonomous driving systems. In this paper, we evaluate the natural and adversarial robustness of various representative models under extensive settings, to fully understand their behaviors influenced by explicit BEV features compared with those without BEV. In addition to the classic settings, we propose a 3D consistent patch attack by applying adversarial patches in the 3D space to guarantee the spatiotemporal consistency, which is more realistic for the scenario of autonomous driving. With substantial experiments, we draw several findings: 1) BEV models tend to be more stable than previous methods under different natural conditions and common corruptions due to the expressive spatial representations; 2) BEV models are more vulnerable to adversarial noises, mainly caused by the redundant BEV features; 3) Camera-LiDAR fusion models have superior performance under different settings with multi-modal inputs, but BEV fusion model is still vulnerable to adversarial noises of both point cloud and image. These findings alert the safety issue in the applications of BEV detectors and could facilitate the development of more robust models.
****
Neural Volumetric Memory for Visual Locomotion Control
Ruihan Yang, Ge Yang, Xiaolong Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1430-1440
Legged robots have the potential to expand the reach of autonomy beyond paved roads. In this work, we consider the difficult problem of locomotion on challenging terrains using a single forward-facing depth camera. Due to the partial observability of the problem, the robot has to rely on past observations to infer the terrain currently beneath it. To solve this problem, we follow the paradigm in computer vision that explicitly models the 3D geometry of the scene and propose Neural Volumetric Memory (NVM), a geometric memory architecture that explicitly accounts for the SE(3) equivariance of the 3D world. NVM aggregates feature volumes from multiple camera views by first bringing them back to the ego-centric frame of the robot. We test the learned visual-locomotion policy on a physical robot and show that our approach, learning legged locomotion with neural volumetric memory, produces performance gains over prior works on challenging terrains. We include ablation studies and show that the representations stored in the neural volumetric memory capture sufficient geometric information to reconstruct the scene. Our project page with videos is https://rchalyang.github.io/NVM/
****
CUF: Continuous Upsampling Filters
Cristina N. Vasconcelos, Cengiz Oztireli, Mark Matthews, Milad Hashemi, Kevin Swersky, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9999-10008
Neural fields have rapidly been adopted for representing 3D signals, but their application to more classical 2D image-processing has been relatively limited. In this paper, we consider one of the most important operations in image processing: upsampling. In deep learning, learnable upsampling layers have extensively been used for single image super-resolution. We propose to parameterize upsampling kernels as neural fields. This parameterization leads to a compact architecture that obtains a 40-fold reduction in the number of parameters when compared with competing arbitrary-scale super-resolution architectures. When upsampling images of size 256x256 we show that our architecture is 2x-10x more efficient than competing arbitrary-scale super-resolution architectures, and more efficient than sub-pixel convolutions when instantiated to a single-scale model. In the general setting, these gains grow polynomially with the square of the target scale. We validate our method on standard benchmarks showing such efficiency gains can be achieved without sacrifices in super-resolution performance.
****
Generalist: Decoupling Natural and Robust Generalization
Hongjun Wang, Yisen Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20554-20563
Deep neural networks obtained by standard training have been constantly plagued by adversarial examples. Although adversarial training demonstrates its capability to defend against adversarial examples, unfortunately, it leads to an inevita

ble drop in the natural generalization. To address the issue, we decouple the natural generalization and the robust generalization from joint training and formulate different training strategies for each one. Specifically, instead of minimizing a global loss on the expectation over these two generalization errors, we propose a bi-expert framework called Generalist where we simultaneously train base learners with task-aware strategies so that they can specialize in their own fields. The parameters of base learners are collected and combined to form a global learner at intervals during the training process. The global learner is then distributed to the base learners as initialized parameters for continued training. Theoretically, we prove that the risks of Generalist will get lower once the base learners are well trained. Extensive experiments verify the applicability of Generalist to achieve high accuracy on natural examples while maintaining considerable robustness to adversarial ones. Code is available at https://github.com/PKU-ML/Generalist.

****

Propagate and Calibrate: Real-Time Passive Non-Line-of-Sight Tracking

Yihao Wang, Zhigang Wang, Bin Zhao, Dong Wang, Mulin Chen, Xuelong Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 972-981

Non-line-of-sight (NLOS) tracking has drawn increasing attention in recent years, due to its ability to detect object motion out of sight. Most previous works on NLOS tracking rely on active illumination, e.g., laser, and suffer from high cost and elaborate experimental conditions. Besides, these techniques are still far from practical application due to oversimplified settings. In contrast, we propose a purely passive method to track a person walking in an invisible room by only observing a relay wall, which is more in line with real application scenarios, e.g., security. To excavate imperceptible changes in videos of the relay wall, we introduce difference frames as an essential carrier of temporal-local motion messages. In addition, we propose PAC-Net, which consists of alternating propagation and calibration, making it capable of leveraging both dynamic and static messages on a frame-level granularity. To evaluate the proposed method, we build and publish the first dynamic passive NLOS tracking dataset, NLOS-Track, which fills the vacuum of realistic NLOS datasets. NLOS-Track contains thousands of NLOS video clips and corresponding trajectories. Both real-shot and synthetic data are included. Our codes and dataset are available at https://againstentropy.github.io/NLOS-Track/.

****

Learning Decorrelated Representations Efficiently Using Fast Fourier Transform

Yutaro Shigeto, Masashi Shimbo, Yuya Yoshikawa, Akikazu Takeuchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2052-2060

Barlow Twins and VICReg are self-supervised representation learning models that use regularizers to decorrelate features. Although these models are as effective as conventional representation learning models, their training can be computationally demanding if the dimension d of the projected embeddings is high. As the regularizers are defined in terms of individual elements of a cross-correlation or covariance matrix, computing the loss for n samples takes O(n d^2) time. In this paper, we propose a relaxed decorrelating regularizer that can be computed in O(n d log d) time by Fast Fourier Transform. We also propose an inexpensive technique to mitigate undesirable local minima that develop with the relaxation. The proposed regularizer exhibits accuracy comparable to that of existing regularizers in downstream tasks, whereas their training requires less memory and is faster for large d. The source code is available.

****

Quantitative Manipulation of Custom Attributes on 3D-Aware Image Synthesis

Hoseok Do, EunKyung Yoo, Taehyeong Kim, Chul Lee, Jin Young Choi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8529-8538

While 3D-based GAN techniques have been successfully applied to render photo-realistic 3D images with a variety of attributes while preserving view consistency,

there has been little research on how to fine-control 3D images without limitin
g to a specific category of objects of their properties. To fill such research g
ap, we propose a novel image manipulation model of 3D-based GAN representations
for a fine-grained control of specific custom attributes. By extending the lates
t 3D-based GAN models (e.g., EG3D), our user-friendly quantitative manipulation
model enables a fine yet normalized control of 3D manipulation of multi-attribut
e quantities while achieving view consistency. We validate the effectiveness of
our proposed technique both qualitatively and quantitatively through various exp
eriments.
****
Explicit Visual Prompting for Low-Level Structure Segmentations
Weihuang Liu, Xi Shen, Chi-Man Pun, Xiaodong Cun; Proceedings of the IEEE/CVF Co
nference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19434-1944
5
We consider the generic problem of detecting low-level structures in images, whi
ch includes segmenting the manipulated parts, identifying out-of-focus pixels, s
eparating shadow regions, and detecting concealed objects. Whereas each such top
ic has been typically addressed with a domain-specific solution, we show that a
unified approach performs well across all of them. We take inspiration from the
widely-used pre-training and then prompt tuning protocols in NLP and propose a n
ew visual prompting model, named Explicit Visual Prompting (EVP). Different from
the previous visual prompting which is typically a dataset-level implicit embed
ding, our key insight is to enforce the tunable parameters focusing on the expli
cit visual content from each individual image, i.e., the features from frozen pa
tch embeddings and the input's high-frequency components. The proposed EVP signi
ficantly outperforms other parameter-efficient tuning protocols under the same a
mount of tunable parameters (5.7% extra trainable parameters of each task). EVP
also achieves state-of-the-art performances on diverse low-level structure segme
ntation tasks compared to task-specific solutions. Our code is available at: htt
ps://github.com/NiFangBaAGe/Explicit-Visual-Prompt.
****
HOTNAS: Hierarchical Optimal Transport for Neural Architecture Search
Jiechao Yang, Yong Liu, Hongteng Xu; Proceedings of the IEEE/CVF Conference on C
omputer Vision and Pattern Recognition (CVPR), 2023, pp. 11990-12000
Instead of searching the entire network directly, current NAS approaches increas
ingly search for multiple relatively small cells to reduce search costs. A major
challenge is to jointly measure the similarity of cell micro-architectures and
the difference in macro-architectures between different cell-based networks. Rec
ently, optimal transport (OT) has been successfully applied to NAS as it can cap
ture the operational and structural similarity across various networks. However,
existing OT-based NAS methods either ignore the cell similarity or focus solely
on searching for a single cell architecture. To address these issues, we propos
e a hierarchical optimal transport metric called HOTNN for measuring the similar
ity of different networks. In HOTNN, the cell-level similarity computes the OT d
istance between cells in various networks by considering the similarity of each
node and the differences in the information flow costs between node pairs within
each cell in terms of operational and structural information. The network-level
similarity calculates OT distance between networks by considering both the cell
-level similarity and the variation in the global position of each cell within t
heir respective networks. We then explore HOTNN in a Bayesian optimization frame
work called HOTNAS, and demonstrate its efficacy in diverse tasks. Extensive exp
eriments demonstrate that HOTNAS can discover network architectures with better
performance in multiple modular cell-based search spaces.
****
Two-Shot Video Object Segmentation
Kun Yan, Xiao Li, Fangyun Wei, Jinglu Wang, Chenbin Zhang, Ping Wang, Yan Lu; Pr
oceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR), 2023, pp. 2257-2267
Previous works on video object segmentation (VOS) are trained on densely annotat
ed videos. Nevertheless, acquiring annotations in pixel level is expensive and t

ime-consuming. In this work, we demonstrate the feasibility of training a satisf actory VOS model on sparsely annotated videos--we merely require two labeled fra mes per training video while the performance is sustained. We term this novel tr aining paradigm as two-shot video object segmentation, or two-shot VOS for short . The underlying idea is to generate pseudo labels for unlabeled frames during t raining and to optimize the model on the combination of labeled and pseudo-label ed data. Our approach is extremely simple and can be applied to a majority of ex isting frameworks. We first pre-train a VOS model on sparsely annotated videos i n a semi-supervised manner, with the first frame always being a labeled one. The n, we adopt the pre-trained VOS model to generate pseudo labels for all unlabele d frames, which are subsequently stored in a pseudo-label bank. Finally, we retr ain a VOS model on both labeled and pseudo-labeled data without any restrictions on the first frame. For the first time, we present a general way to train VOS m odels on two-shot VOS datasets. By using 7.3% and 2.9% labeled data of YouTube-V OS and DAVIS benchmarks, our approach achieves comparable results in contrast to the counterparts trained on fully labeled set. Code and models are available at https://github.com/yk-pku/Two-shot-Video-Object-Segmentation.
****
Neural Fields Meet Explicit Geometric Representations for Inverse Rendering of U rban Scenes
Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgre n, Zan Gojcic, Wenzheng Chen, Sanja Fidler; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8370-8380
Reconstruction and intrinsic decomposition of scenes from captured imagery would enable many applications such as relighting and virtual object insertion. Recen t NeRF based methods achieve impressive fidelity of 3D reconstruction, but bake the lighting and shadows into the radiance field, while mesh-based methods that facilitate intrinsic decomposition through differentiable rendering have not yet scaled to the complexity and scale of outdoor scenes. We present a novel invers e rendering framework for large urban scenes capable of jointly reconstructing t he scene geometry, spatially-varying materials, and HDR lighting from a set of p osed RGB images with optional depth. Specifically, we use a neural field to acco unt for the primary rays, and use an explicit mesh (reconstructed from the under lying neural field) for modeling secondary rays that produce higher-order lighti ng effects such as cast shadows. By faithfully disentangling complex geometry an d materials from lighting effects, our method enables photorealistic relighting with specular and shadow effects on several outdoor datasets. Moreover, it suppo rts physics-based scene manipulations such as virtual object insertion with ray-traced shadow casting.
****
Practical Network Acceleration With Tiny Sets
Guo-Hua Wang, Jianxin Wu; Proceedings of the IEEE/CVF Conference on Computer Vis ion and Pattern Recognition (CVPR), 2023, pp. 20331-20340
Due to data privacy issues, accelerating networks with tiny training sets has be come a critical need in practice. Previous methods mainly adopt filter-level pru ning to accelerate networks with scarce training samples. In this paper, we reve al that dropping blocks is a fundamentally superior approach in this scenario. I t enjoys a higher acceleration ratio and results in a better latency-accuracy pe rformance under the few-shot setting. To choose which blocks to drop, we propose a new concept namely recoverability to measure the difficulty of recovering the compressed network. Our recoverability is efficient and effective for choosing which blocks to drop. Finally, we propose an algorithm named PRACTISE to acceler ate networks using only tiny sets of training images. PRACTISE outperforms previ ous methods by a significant margin. For 22% latency reduction, PRACTISE surpass es previous methods by on average 7% on ImageNet-1k. It also enjoys high general ization ability, working well under data-free or out-of-domain data settings, to o. Our code is at https://github.com/DoctorKey/Practise.
****
NeRF-RPN: A General Framework for Object Detection in NeRFs
Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, Chi-Keung Tang; Proceedings of

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23528-23538

This paper presents the first significant object detection framework, NeRF-RPN, which directly operates on NeRF. Given a pre-trained NeRF model, NeRF-RPN aims to detect all bounding boxes of objects in a scene. By exploiting a novel voxel representation that incorporates multi-scale 3D neural volumetric features, we demonstrate it is possible to regress the 3D bounding boxes of objects in NeRF directly without rendering the NeRF at any viewpoint. NeRF-RPN is a general framework and can be applied to detect objects without class labels. We experimented NeRF-RPN with various backbone architectures, RPN head designs, and loss functions. All of them can be trained in an end-to-end manner to estimate high quality 3D bounding boxes. To facilitate future research in object detection for NeRF, we built a new benchmark dataset which consists of both synthetic and real-world data with careful labeling and clean up. Code and dataset are available at https://github.com/lyclyc52/NeRF_RPN.

****

Cross-Image-Attention for Conditional Embeddings in Deep Metric Learning

Dmytro Kotovenko, Pingchuan Ma, Timo Milbich, Björn Ommer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11070-11081

Learning compact image embeddings that yield semantic similarities between images and that generalize to unseen test classes, is at the core of deep metric learning (DML). Finding a mapping from a rich, localized image feature map onto a compact embedding vector is challenging: Although similarity emerges between tuples of images, DML approaches marginalize out information in an individual image before considering another image to which similarity is to be computed. Instead, we propose during training to condition the embedding of an image on the image we want to compare it to. Rather than embedding by a simple pooling as in standard DML, we use cross-attention so that one image can identify relevant features in the other image. Consequently, the attention mechanism establishes a hierarchy of conditional embeddings that gradually incorporates information about the tuple to steer the representation of an individual image. The cross-attention layers bridge the gap between the original unconditional embedding and the final similarity and allow backpropagtion to update encodings more directly than through a lossy pooling layer. At test time we use the resulting improved unconditional embeddings, thus requiring no additional parameters or computational overhead. Experiments on established DML benchmarks show that our cross-attention conditional embedding during training improves the underlying standard DML pipeline significantly so that it outperforms the state-of-the-art.

****

Masked Wavelet Representation for Compact Neural Radiance Fields

Daniel Rho, Byeonghyeon Lee, Seungtae Nam, Joo Chan Lee, Jong Hwan Ko, Eunbyung Park; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20680-20690

Neural radiance fields (NeRF) have demonstrated the potential of coordinate-based neural representation (neural fields or implicit neural representation) in neural rendering. However, using a multi-layer perceptron (MLP) to represent a 3D scene or object requires enormous computational resources and time. There have been recent studies on how to reduce these computational inefficiencies by using additional data structures, such as grids or trees. Despite the promising performance, the explicit data structure necessitates a substantial amount of memory. In this work, we present a method to reduce the size without compromising the advantages of having additional data structures. In detail, we propose using the wavelet transform on grid-based neural fields. Grid-based neural fields are for fast convergence, and the wavelet transform, whose efficiency has been demonstrated in high-performance standard codecs, is to improve the parameter efficiency of grids. Furthermore, in order to achieve a higher sparsity of grid coefficients while maintaining reconstruction quality, we present a novel trainable masking approach. Experimental results demonstrate that non-spatial grid coefficients, such as wavelet coefficients, are capable of attaining a higher level of sparsity

than spatial grid coefficients, resulting in a more compact representation. With our proposed mask and compression pipeline, we achieved state-of-the-art performance within a memory budget of 2 MB. Our code is available at https://github.com/daniel03c1/masked_wavelet_nerf.
****
PiMAE: Point Cloud and Image Interactive Masked Autoencoders for 3D Object Detection
Anthony Chen, Kevin Zhang, Renrui Zhang, Zihan Wang, Yuheng Lu, Yandong Guo, Shanghang Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5291-5301
Masked Autoencoders learn strong visual representations and achieve state-of-the-art results in several independent modalities, yet very few works have addressed their capabilities in multi-modality settings. In this work, we focus on point cloud and RGB image data, two modalities that are often presented together in the real world and explore their meaningful interactions. To improve upon the cross-modal synergy in existing works, we propose PiMAE, a self-supervised pre-training framework that promotes 3D and 2D interaction through three aspects. Specifically, we first notice the importance of masking strategies between the two sources and utilize a projection module to complementarily align the mask and visible tokens of the two modalities. Then, we utilize a well-crafted two-branch MAE pipeline with a novel shared decoder to promote cross-modality interaction in the mask tokens. Finally, we design a unique cross-modal reconstruction module to enhance representation learning for both modalities. Through extensive experiments performed on large-scale RGB-D scene understanding benchmarks (SUN RGB-D and ScannetV2), we discover it is nontrivial to interactively learn point-image features, where we greatly improve multiple 3D detectors, 2D detectors and few-shot classifiers by 2.9%, 6.7%, and 2.4%, respectively. Code is available at https://github.com/BLVLab/PiMAE.
****
ObjectStitch: Object Compositing With Diffusion Model
Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, Daniel Aliaga; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18310-18319
Object compositing based on 2D images is a challenging problem since it typically involves multiple processing stages such as color harmonization, geometry correction and shadow generation to generate realistic results. Furthermore, annotating training data pairs for compositing requires substantial manual effort from professionals, and is hardly scalable. Thus, with the recent advances in generative models, in this work, we propose a self-supervised framework for object compositing by leveraging the power of conditional diffusion models. Our framework can hollistically address the object compositing task in a unified model, transforming the viewpoint, geometry, color and shadow of the generated object while requiring no manual labeling. To preserve the input object's characteristics, we introduce a content adaptor that helps to maintain categorical semantics and object appearance. A data augmentation method is further adopted to improve the fidelity of the generator. Our method outperforms relevant baselines in both realism and faithfulness of the synthesized result images in a user study on various real-world images.
****
High-Fidelity 3D GAN Inversion by Pseudo-Multi-View Optimization
Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, Qifeng Chen; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 321-331
We present a high-fidelity 3D generative adversarial network (GAN) inversion framework that can synthesize photo-realistic novel views while preserving specific details of the input image. High-fidelity 3D GAN inversion is inherently challenging due to the geometry-texture trade-off, where overfitting to a single view input image often damages the estimated geometry during the latent optimization. To solve this challenge, we propose a novel pipeline that builds on the pseudo-multi-view estimation with visibility analysis. We keep the original textures fo

r the visible parts and utilize generative priors for the occluded parts. Extensive experiments show that our approach achieves advantageous reconstruction and novel view synthesis quality over prior work, even for images with out-of-distribution textures. The proposed pipeline also enables image attribute editing with the inverted latent code and 3D-aware texture modification. Our approach enables high-fidelity 3D rendering from a single image, which is promising for various applications of AI-generated 3D content. The source code is at https://github.com/jiaxinxie97/HFGI3D/.
****
Anchor3DLane: Learning To Regress 3D Anchors for Monocular 3D Lane Detection
Shaofei Huang, Zhenwei Shen, Zehao Huang, Zi-han Ding, Jiao Dai, Jizhong Han, Naiyan Wang, Si Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17451-17460
Monocular 3D lane detection is a challenging task due to its lack of depth information. A popular solution is to first transform the front-viewed (FV) images or features into the bird-eye-view (BEV) space with inverse perspective mapping (IPM) and detect lanes from BEV features. However, the reliance of IPM on flat ground assumption and loss of context information make it inaccurate to restore 3D information from BEV representations. An attempt has been made to get rid of BEV and predict 3D lanes from FV representations directly, while it still underperforms other BEV-based methods given its lack of structured representation for 3D lanes. In this paper, we define 3D lane anchors in the 3D space and propose a BEV-free method named Anchor3DLane to predict 3D lanes directly from FV representations. 3D lane anchors are projected to the FV features to extract their features which contain both good structural and context information to make accurate predictions. In addition, we also develop a global optimization method that makes use of the equal-width property between lanes to reduce the lateral error of predictions. Extensive experiments on three popular 3D lane detection benchmarks show that our Anchor3DLane outperforms previous BEV-based methods and achieves state-of-the-art performances. The code is available at: https://github.com/tusen-ai/Anchor3DLane.
****
Class-Balancing Diffusion Models
Yiming Qin, Huangjie Zheng, Jiangchao Yao, Mingyuan Zhou, Ya Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18434-18443
Diffusion-based models have shown the merits of generating high-quality visual data while preserving better diversity in recent studies. However, such observation is only justified with curated data distribution, where the data samples are nicely pre-processed to be uniformly distributed in terms of their labels. In practice, a long-tailed data distribution appears more common and how diffusion models perform on such class-imbalanced data remains unknown. In this work, we first investigate this problem and observe significant degradation in both diversity and fidelity when the diffusion model is trained on datasets with class-imbalanced distributions. Especially in tail classes, the generations largely lose diversity and we observe severe mode-collapse issues. To tackle this problem, we set from the hypothesis that the data distribution is not class-balanced, and propose Class-Balancing Diffusion Models (CBDM) that are trained with a distribution adjustment regularizer as a solution. Experiments show that images generated by CBDM exhibit higher diversity and quality in both quantitative and qualitative ways. Our method benchmarked the generation results on CIFAR100/CIFAR100LT dataset and shows outstanding performance on the downstream recognition task.
****
AstroNet: When Astrocyte Meets Artificial Neural Network
Mengqiao Han, Liyuan Pan, Xiabi Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20258-20268
Network structure learning aims to optimize network architectures and make them more efficient without compromising performance. In this paper, we first study the astrocytes, a new mechanism to regulate connections in the classic M-P neuron. Then, with the astrocytes, we propose an AstroNet that can adaptively optimize

neuron connections and therefore achieves structure learning to achieve higher accuracy and efficiency. AstroNet is based on our built Astrocyte-Neuron model, with a temporal regulation mechanism and a global connection mechanism, which is inspired by the bidirectional communication property of astrocytes. With the model, the proposed AstroNet uses a neural network (NN) for performing tasks, and an astrocyte network (AN) to continuously optimize the connections of NN, i.e., assigning weight to the neuron units in the NN adaptively. Experiments on the classification task demonstrate that our AstroNet can efficiently optimize the network structure while achieving state-of-the-art (SOTA) accuracy.
****

Feature Alignment and Uniformity for Test Time Adaptation
Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, Rui Li; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20050-20060
Test time adaptation (TTA) aims to adapt deep neural networks when receiving out of distribution test domain samples. In this setting, the model can only access online unlabeled test samples and pre-trained models on the training domains. We first address TTA as a feature revision problem due to the domain gap between source domains and target domains. After that, we follow the two measurements alignment and uniformity to discuss the test time feature revision. For test time feature uniformity, we propose a test time self-distillation strategy to guarantee the consistency of uniformity between representations of the current batch and all the previous batches. For test time feature alignment, we propose a memorized spatial local clustering strategy to align the representations among the neighborhood samples for the upcoming batch. To deal with the common noisy label problem, we propound the entropy and consistency filters to select and drop the possible noisy labels. To prove the scalability and efficacy of our method, we conduct experiments on four domain generalization benchmarks and four medical image segmentation tasks with various backbones. Experiment results show that our method not only improves baseline stably but also outperforms existing state-of-the-art test time adaptation methods.
****

Balanced Product of Calibrated Experts for Long-Tailed Recognition
Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, Marco Kuhlmann; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19967-19977
Many real-world recognition problems are characterized by long-tailed label distributions. These distributions make representation learning highly challenging due to limited generalization over the tail classes. If the test distribution differs from the training distribution, e.g. uniform versus long-tailed, the problem of the distribution shift needs to be addressed. A recent line of work proposes learning multiple diverse experts to tackle this issue. Ensemble diversity is encouraged by various techniques, e.g. by specializing different experts in the head and the tail classes. In this work, we take an analytical approach and extend the notion of logit adjustment to ensembles to form a Balanced Product of Experts (BalPoE). BalPoE combines a family of experts with different test-time target distributions, generalizing several previous approaches. We show how to properly define these distributions and combine the experts in order to achieve unbiased predictions, by proving that the ensemble is Fisher-consistent for minimizing the balanced error. Our theoretical analysis shows that our balanced ensemble requires calibrated experts, which we achieve in practice using mixup. We conduct extensive experiments and our method obtains new state-of-the-art results on three long-tailed datasets: CIFAR-100-LT, ImageNet-LT, and iNaturalist-2018. Our code is available at https://github.com/emasa/BalPoE-CalibratedLT.
****

Single Image Backdoor Inversion via Robust Smoothed Classifiers
Mingjie Sun, Zico Kolter; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8113-8122
Backdoor inversion, the process of finding a backdoor trigger inserted into a machine learning model, has become the pillar of many backdoor detection and defen

se methods. Previous works on backdoor inversion often recover the backdoor through an optimization process to flip a support set of clean images into the target class. However, it is rarely studied and understood how large this support set should be to recover a successful backdoor. In this work, we show that one can reliably recover the backdoor trigger with as few as a single image. Specifically, we propose the SmoothInv method, which first constructs a robust smoothed version of the backdoored classifier and then performs guided image synthesis towards the target class to reveal the backdoor pattern. SmoothInv requires neither an explicit modeling of the backdoor via a mask variable, nor any complex regularization schemes, which has become the standard practice in backdoor inversion methods. We perform both quantitaive and qualitative study on backdoored classifiers from previous published backdoor attacks. We demonstrate that compared to existing methods, SmoothInv is able to recover successful backdoors from single images, while maintaining high fidelity to the original backdoor. We also show how we identify the target backdoored class from the backdoored classifier. Last, we propose and analyze two countermeasures to our approach and show that SmoothInv remains robust in the face of an adaptive attacker. Our code is available at https://github.com/locuslab/smoothinv.
****
PanoSwin: A Pano-Style Swin Transformer for Panorama Understanding
Zhixin Ling, Zhen Xing, Xiangdong Zhou, Manliang Cao, Guichun Zhou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17755-17764
In panorama understanding, the widely used equirectangular projection (ERP) entails boundary discontinuity and spatial distortion. It severely deteriorates the conventional CNNs and vision Transformers on panoramas. In this paper, we propose a simple yet effective architecture named PanoSwin to learn panorama representations with ERP. To deal with the challenges brought by equirectangular projection, we explore a pano-style shift windowing scheme and novel pitch attention to address the boundary discontinuity and the spatial distortion, respectively. Besides, based on spherical distance and Cartesian coordinates, we adapt absolute positional encodings and relative positional biases for panoramas to enhance panoramic geometry information. Realizing that planar image understanding might share some common knowledge with panorama understanding, we devise a novel two-stage learning framework to facilitate knowledge transfer from the planar images to panoramas. We conduct experiments against the state-of-the-art on various panoramic tasks, i.e., panoramic object detection, panoramic classification, and panoramic layout estimation. The experimental results demonstrate the effectiveness of PanoSwin in panorama understanding.
****
Parameter Efficient Local Implicit Image Function Network for Face Segmentation
Mausoom Sarkar, Nikitha SR, Mayur Hemani, Rishabh Jain, Balaji Krishnamurthy; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20970-20980
Face parsing is defined as the per-pixel labeling of images containing human faces. The labels are defined to identify key facial regions like eyes, lips, nose, hair, etc. In this work, we make use of the structural consistency of the human face to propose a lightweight face-parsing method using a Local Implicit Function network, FP-LIIF. We propose a simple architecture having a convolutional encoder and a pixel MLP decoder that uses 1/26th number of parameters compared to the state-of-the-art models and yet matches or outperforms state-of-the-art models on multiple datasets, like CelebAMask-HQ and LaPa. We do not use any pretraining, and compared to other works, our network can also generate segmentation at different resolutions without any changes in the input resolution. This work enables the use of facial segmentation on low-compute or low-bandwidth devices because of its higher FPS and smaller model size.
****
A Hierarchical Representation Network for Accurate and Detailed Face Reconstruction From In-the-Wild Images
Biwen Lei, Jianqiang Ren, Mengyang Feng, Miaomiao Cui, Xuansong Xie; Proceedings

 of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2
023, pp. 394-403
Limited by the nature of the low-dimensional representational capacity of 3DMM,
most of the 3DMM-based face reconstruction (FR) methods fail to recover high-fre
quency facial details, such as wrinkles, dimples, etc. Some attempt to solve the
 problem by introducing detail maps or non-linear operations, however, the resul
ts are still not vivid. To this end, we in this paper present a novel hierarchic
al representation network (HRN) to achieve accurate and detailed face reconstruc
tion from a single image. Specifically, we implement the geometry disentanglemen
t and introduce the hierarchical representation to fulfill detailed face modelin
g. Meanwhile, 3D priors of facial details are incorporated to enhance the accura
cy and authenticity of the reconstruction results. We also propose a de-retouchi
ng module to achieve better decoupling of the geometry and appearance. It is not
eworthy that our framework can be extended to a multi-view fashion by considerin
g detail consistency of different views. Extensive experiments on two single-vie
w and two multi-view FR benchmarks demonstrate that our method outperforms the e
xisting methods in both reconstruction accuracy and visual effects. Finally, we
introduce a high-quality 3D face dataset FaceHD-100 to boost the research of hig
h-fidelity face reconstruction. The project homepage is at https://younglbw.gith
ub.io/HRN-homepage/.
****
PersonNeRF: Personalized Reconstruction From Photo Collections
Chung-Yi Weng, Pratul P. Srinivasan, Brian Curless, Ira Kemelmacher-Shlizerman;
Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognitio
n (CVPR), 2023, pp. 524-533
We present PersonNeRF, a method that takes a collection of photos of a subject (
e.g., Roger Federer) captured across multiple years with arbitrary body poses an
d appearances, and enables rendering the subject with arbitrary novel combinatio
ns of viewpoint, body pose, and appearance. PersonNeRF builds a customized neura
l volumetric 3D model of the subject that is able to render an entire space span
ned by camera viewpoint, body pose, and appearance. A central challenge in this
task is dealing with sparse observations; a given body pose is likely only obser
ved by a single viewpoint with a single appearance, and a given appearance is on
ly observed under a handful of different body poses. We address this issue by re
covering a canonical T-pose neural volumetric representation of the subject that
 allows for changing appearance across different observations, but uses a shared
 pose-dependent motion field across all observations. We demonstrate that this a
pproach, along with regularization of the recovered volumetric geometry to encou
rage smoothness, is able to recover a model that renders compelling images from
novel combinations of viewpoint, pose, and appearance from these challenging uns
tructured photo collections, outperforming prior work for free-viewpoint human r
endering.
****
Enhanced Multimodal Representation Learning With Cross-Modal KD
Mengxi Chen, Linyu Xing, Yu Wang, Ya Zhang; Proceedings of the IEEE/CVF Conferen
ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11766-11775
This paper explores the tasks of leveraging auxiliary modalities which are only
available at training to enhance multimodal representation learning through cros
s-modal Knowledge Distillation (KD). The widely adopted mutual information maxim
ization-based objective leads to a short-cut solution of the weak teacher, i.e.,
 achieving the maximum mutual information by simply making the teacher model as
weak as the student model. To prevent such a weak solution, we introduce an addi
tional objective term, i.e., the mutual information between the teacher and the
auxiliary modality model. Besides, to narrow down the information gap between th
e student and teacher, we further propose to minimize the conditional entropy of
 the teacher given the student. Novel training schemes based on contrastive lear
ning and adversarial learning are designed to optimize the mutual information an
d the conditional entropy, respectively. Experimental results on three popular m
ultimodal benchmark datasets have shown that the proposed method outperforms a r
ange of state-of-the-art approaches for video recognition, video retrieval and e

motion classification.
****
Learning a Depth Covariance Function
Eric Dexheimer, Andrew J. Davison; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13122-13131
We propose learning a depth covariance function with applications to geometric vision tasks. Given RGB images as input, the covariance function can be flexibly used to define priors over depth functions, predictive distributions given observations, and methods for active point selection. We leverage these techniques for a selection of downstream tasks: depth completion, bundle adjustment, and monocular dense visual odometry.
****
Evading DeepFake Detectors via Adversarial Statistical Consistency
Yang Hou, Qing Guo, Yihao Huang, Xiaofei Xie, Lei Ma, Jianjun Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12271-12280
In recent years, as various realistic face forgery techniques known as DeepFake improves by leaps and bounds, more and more DeepFake detection techniques have been proposed. These methods typically rely on detecting statistical differences between natural (i.e., real) and DeepFake-generated images in both spatial and frequency domains. In this work, we propose to explicitly minimize the statistical differences to evade state-of-the-art DeepFake detectors. To this end, we propose a statistical consistency attack (StatAttack) against DeepFake detectors, which contains two main parts. First, we select several statistical-sensitive natural degradations (i.e., exposure, blur, and noise) and add them to the fake images in an adversarial way. Second, we find that the statistical differences between natural and DeepFake images are positively associated with the distribution shifting between the two kinds of images, and we propose to use a distribution-aware loss to guide the optimization of different degradations. As a result, the feature distributions of generated adversarial examples is close to the natural images. Furthermore, we extend the StatAttack to a more powerful version, MStatAttack, where we extend the single-layer degradation to multi-layer degradations sequentially and use the loss to tune the combination weights jointly. Comprehensive experimental results on four spatial-based detectors and two frequency-based detectors with four datasets demonstrate the effectiveness of our proposed attack method in both white-box and black-box settings.
****
Referring Image Matting
Jizhizi Li, Jing Zhang, Dacheng Tao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22448-22457
Different from conventional image matting, which either requires user-defined scribbles/trimap to extract a specific foreground object or directly extracts all the foreground objects in the image indiscriminately, we introduce a new task named Referring Image Matting (RIM) in this paper, which aims to extract the meticulous alpha matte of the specific object that best matches the given natural language description, thus enabling a more natural and simpler instruction for image matting. First, we establish a large-scale challenging dataset RefMatte by designing a comprehensive image composition and expression generation engine to automatically produce high-quality images along with diverse text attributes based on public datasets. RefMatte consists of 230 object categories, 47,500 images, 118,749 expression-region entities, and 474,996 expressions. Additionally, we construct a real-world test set with 100 high-resolution natural images and manually annotate complex phrases to evaluate the out-of-domain generalization abilities of RIM methods. Furthermore, we present a novel baseline method CLIPMat for RIM, including a context-embedded prompt, a text-driven semantic pop-up, and a multi-level details extractor. Extensive experiments on RefMatte in both keyword and expression settings validate the superiority of CLIPMat over representative methods. We hope this work could provide novel insights into image matting and encourage more follow-up studies. The dataset, code and models are available at https://github.com/JizhiziLi/RIM.

****
V2V4Real: A Real-World Large-Scale Dataset for Vehicle-to-Vehicle Cooperative Pe rception
Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, Jiaqi Ma; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13712-13722
Modern perception systems of autonomous vehicles are known to be sensitive to occlusions and lack the capability of long perceiving range. It has been one of the key bottlenecks that prevents Level 5 autonomy. Recent research has demonstrated that the Vehicle-to-Vehicle (V2V) cooperative perception system has great potential to revolutionize the autonomous driving industry. However, the lack of a real-world dataset hinders the progress of this field. To facilitate the development of cooperative perception, we present V2V4Real, the first large-scale real-world multi-modal dataset for V2V perception. The data is collected by two vehicles equipped with multi-modal sensors driving together through diverse scenarios . Our V2V4Real dataset covers a driving area of 410 km, comprising 20K LiDAR frames, 40K RGB frames, 240K annotated 3D bounding boxes for 5 classes, and HDMaps that cover all the driving routes. V2V4Real introduces three perception tasks, including cooperative 3D object detection, cooperative 3D object tracking, and Sim2Real domain adaptation for cooperative perception. We provide comprehensive benchmarks of recent cooperative perception algorithms on three tasks. The V2V4Real dataset can be found at research.seas.ucla.edu/mobility-lab/v2v4real/.
****
RMLVQA: A Margin Loss Approach for Visual Question Answering With Language Biase s
Abhipsa Basu, Sravanti Addepalli, R. Venkatesh Babu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11671-11680
Visual Question Answering models have been shown to suffer from language biases, where the model learns a correlation between the question and the answer, ignoring the image. While early works attempted to use question-only models or data augmentations to reduce this bias, we propose an adaptive margin loss approach having two components. The first component considers the frequency of answers within a question type in the training data, which addresses the concern of the class-imbalance causing the language biases. However, it does not take into account the answering difficulty of the samples, which impacts their learning. We address this through the second component, where instance-specific margins are learnt, allowing the model to distinguish between samples of varying complexity. We introduce a bias-injecting component to our model, and compute the instance-specific margins from the confidence of this component. We combine these with the estimated margins to consider both answer-frequency and task-complexity in the training loss. We show that, while the margin loss is effective for out-of-distribution (ood) data, the bias-injecting component is essential for generalising to in-distribution (id) data. Our proposed approach, Robust Margin Loss for Visual Question Answering (RMLVQA) improves upon the existing state-of-the-art results when compared to augmentation-free methods on benchmark VQA datasets suffering from language biases, while maintaining competitive performance on id data, making our method the most robust one among all comparable methods.
****
NeuralLift-360: Lifting an In-the-Wild 2D Photo to a 3D Object With 360deg Views
Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, Zhangyang Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4479-4489
Virtual reality and augmented reality (XR) bring increasing demand for 3D content generation. However, creating high-quality 3D content requires tedious work from a human expert. In this work, we study the challenging task of lifting a single image to a 3D object and, for the first time, demonstrate the ability to generate a plausible 3D object with 360deg views that corresponds well with the given reference image. By conditioning on the reference image, our model can fulfill

the everlasting curiosity for synthesizing novel views of objects from images. Our technique sheds light on a promising direction of easing the workflows for 3D artists and XR designers. We propose a novel framework, dubbed NeuralLift-360, that utilizes a depth-aware neural radiance representation (NeRF) and learns to craft the scene guided by denoising diffusion models. By introducing a ranking loss, our NeuralLift-360 can be guided with rough depth estimation in the wild. We also adopt a CLIP-guided sampling strategy for the diffusion prior to provide coherent guidance. Extensive experiments demonstrate that our NeuralLift-360 significantly outperforms existing state-of-the-art baselines. Project page: https://vita-group.github.io/NeuralLift-360/
****
ViP3D: End-to-End Visual Trajectory Prediction via 3D Agent Queries
Junru Gu, Chenxu Hu, Tianyuan Zhang, Xuanyao Chen, Yilun Wang, Yue Wang, Hang Zhao; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5496-5506
Perception and prediction are two separate modules in the existing autonomous driving systems. They interact with each other via hand-picked features such as agent bounding boxes and trajectories. Due to this separation, prediction, as a downstream module, only receives limited information from the perception module. To make matters worse, errors from the perception modules can propagate and accumulate, adversely affecting the prediction results. In this work, we propose ViP3D, a query-based visual trajectory prediction pipeline that exploits rich information from raw videos to directly predict future trajectories of agents in a scene. ViP3D employs sparse agent queries to detect, track, and predict throughout the pipeline, making it the first fully differentiable vision-based trajectory prediction approach. Instead of using historical feature maps and trajectories, useful information from previous timestamps is encoded in agent queries, which makes ViP3D a concise streaming prediction method. Furthermore, extensive experimental results on the nuScenes dataset show the strong vision-based prediction performance of ViP3D over traditional pipelines and previous end-to-end models.
****
Modality-Invariant Visual Odometry for Embodied Vision
Marius Memmel, Roman Bachmann, Amir Zamir; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21549-21559
Effectively localizing an agent in a realistic, noisy setting is crucial for many embodied vision tasks. Visual Odometry (VO) is a practical substitute for unreliable GPS and compass sensors, especially in indoor environments. While SLAM-based methods show a solid performance without large data requirements, they are less flexible and robust w.r.t. to noise and changes in the sensor suite compared to learning-based approaches. Recent deep VO models, however, limit themselves to a fixed set of input modalities, e.g., RGB and depth, while training on millions of samples. When sensors fail, sensor suites change, or modalities are intentionally looped out due to available resources, e.g., power consumption, the models fail catastrophically. Furthermore, training these models from scratch is even more expensive without simulator access or suitable existing models that can be fine-tuned. While such scenarios get mostly ignored in simulation, they commonly hinder a model's reusability in real-world applications. We propose a Transformer-based modality-invariant VO approach that can deal with diverse or changing sensor suites of navigation agents. Our model outperforms previous methods while training on only a fraction of the data. We hope this method opens the door to a broader range of real-world applications that can benefit from flexible and learned VO models.
****
What You Can Reconstruct From a Shadow
Ruoshi Liu, Sachit Menon, Chengzhi Mao, Dennis Park, Simon Stent, Carl Vondrick; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17059-17068
3D reconstruction is a fundamental problem in computer vision, and the task is especially challenging when the object to reconstruct is partially or fully occluded. We introduce a method that uses the shadows cast by an unobserved object in

order to infer the possible 3D volumes under occlusion. We create a differentiable image formation model that allows us to jointly infer the 3D shape of an object, its pose, and the position of a light source. Since the approach is end-to-end differentiable, we are able to integrate learned priors of object geometry in order to generate realistic 3D shapes of different object categories. Experiments and visualizations show that the method is able to generate multiple possible solutions that are consistent with the observation of the shadow. Our approach works even when the position of the light source and object pose are both unknown. Our approach is also robust to real-world images where ground-truth shadow mask is unknown.
****
Adaptive Sparse Convolutional Networks With Global Context Enhancement for Faster Object Detection on Drone Images
Bowei Du, Yecheng Huang, Jiaxin Chen, Di Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13435-13444
Object detection on drone images with low-latency is an important but challenging task on the resource-constrained unmanned aerial vehicle (UAV) platform. This paper investigates optimizing the detection head based on the sparse convolution, which proves effective in balancing the accuracy and efficiency. Nevertheless, it suffers from inadequate integration of contextual information of tiny objects as well as clumsy control of the mask ratio in the presence of foreground with varying scales. To address the issues above, we propose a novel global context-enhanced adaptive sparse convolutional network (CEASC). It first develops a context-enhanced group normalization (CE-GN) layer, by replacing the statistics based on sparsely sampled features with the global contextual ones, and then designs an adaptive multi-layer masking strategy to generate optimal mask ratios at distinct scales for compact foreground coverage, promoting both the accuracy and efficiency. Extensive experimental results on two major benchmarks, i.e. VisDrone and UAVDT, demonstrate that CEASC remarkably reduces the GFLOPs and accelerates the inference procedure when plugging into the typical state-of-the-art detection frameworks (e.g. RetinaNet and GFL V1) with competitive performance. Code is available at https://github.com/Cuogeihong/CEASC.
****
LidarGait: Benchmarking 3D Gait Recognition With Point Clouds
Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, Shiqi Yu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1054-1063
Video-based gait recognition has achieved impressive results in constrained scenarios. However, visual cameras neglect human 3D structure information, which limits the feasibility of gait recognition in the 3D wild world. Instead of extracting gait features from images, this work explores precise 3D gait features from point clouds and proposes a simple yet efficient 3D gait recognition framework, termed LidarGait. Our proposed approach projects sparse point clouds into depth maps to learn the representations with 3D geometry information, which outperforms existing point-wise and camera-based methods by a significant margin. Due to the lack of point cloud datasets, we build the first large-scale LiDAR-based gait recognition dataset, SUSTech1K, collected by a LiDAR sensor and an RGB camera. The dataset contains 25,239 sequences from 1,050 subjects and covers many variations, including visibility, views, occlusions, clothing, carrying, and scenes. Extensive experiments show that (1) 3D structure information serves as a significant feature for gait recognition. (2) LidarGait outperforms existing point-based and silhouette-based methods by a significant margin, while it also offers stable cross-view results. (3) The LiDAR sensor is superior to the RGB camera for gait recognition in the outdoor environment. The source code and dataset have been made available at https://lidargait.github.io.
****
Command-Driven Articulated Object Understanding and Manipulation
Ruihang Chu, Zhengzhe Liu, Xiaoqing Ye, Xiao Tan, Xiaojuan Qi, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 8813-8823

We present Cart, a new approach towards articulated-object manipulations by human commands. Beyond the existing work that focuses on inferring articulation structures, we further support manipulating articulated shapes to align them subject to simple command templates. The key of Cart is to utilize the prediction of object structures to connect visual observations with user commands for effective manipulations. It is achieved by encoding command messages for motion prediction and a test-time adaptation to adjust the amount of movement from only command supervision. For a rich variety of object categories, Cart can accurately manipulate object shapes and outperform the state-of-the-art approaches in understanding the inherent articulation structures. Also, it can well generalize to unseen object categories and real-world objects. We hope Cart could open new directions for instructing machines to operate articulated objects.
****
D2Former: Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-Based Transformers
Jianfeng He, Yuan Gao, Tianzhu Zhang, Zhe Zhang, Feng Wu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2904-2914
Establishing pixel-level matches between image pairs is vital for a variety of computer vision applications. However, achieving robust image matching remains challenging because CNN extracted descriptors usually lack discriminative ability in texture-less regions and keypoint detectors are only good at identifying keypoints with a specific level of structure. To deal with these issues, a novel image matching method is proposed by Jointly Learning Hierarchical Detectors and Contextual Descriptors via Agent-based Transformers (D2Former), including a contextual feature descriptor learning (CFDL) module and a hierarchical keypoint detector learning (HKDL) module. The proposed D2Former enjoys several merits. First, the proposed CFDL module can model long-range contexts efficiently and effectively with the aid of designed descriptor agents. Second, the HKDL module can generate keypoint detectors in a hierarchical way, which is helpful for detecting keypoints with diverse levels of structures. Extensive experimental results on four challenging benchmarks show that our proposed method significantly outperforms state-of-the-art image matching methods.
****
ConStruct-VL: Data-Free Continual Structured VL Concepts Learning
James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, Leonid Karlinsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14994-15004
Recently, large-scale pre-trained Vision-and-Language (VL) foundation models have demonstrated remarkable capabilities in many zero-shot downstream tasks, achieving competitive results for recognizing objects defined by as little as short text prompts. However, it has also been shown that VL models are still brittle in Structured VL Concept (SVLC) reasoning, such as the ability to recognize object attributes, states, and inter-object relations. This leads to reasoning mistakes, which need to be corrected as they occur by teaching VL models the missing SVLC skills; often this must be done using private data where the issue was found, which naturally leads to a data-free continual (no task-id) VL learning setting. In this work, we introduce the first Continual Data-Free Structured VL Concepts Learning (ConStruct-VL) benchmark and show it is challenging for many existing data-free CL strategies. We, therefore, propose a data-free method comprised of a new approach of Adversarial Pseudo-Replay (APR) which generates adversarial reminders of past tasks from past task models. To use this method efficiently, we also propose a continual parameter-efficient Layered-LoRA (LaLo) neural architecture allowing no-memory-cost access to all past models at train time. We show this approach outperforms all data-free methods by as much as 7% while even matching some levels of experience-replay (prohibitive for applications where data-privacy must be preserved). Our code is publicly available at https://github.com/jamessealesmith/ConStruct-VL
****

Lite DETR: An Interleaved Multi-Scale Encoder for Efficient DETR
Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, Lionel M. Ni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18558-18567
Recent DEtection TRansformer-based (DETR) models have obtained remarkable performance. Its success cannot be achieved without the re-introduction of multi-scale feature fusion in the encoder. However, the excessively increased tokens in multi-scale features, especially for about 75% of low-level features, are quite computationally inefficient, which hinders real applications of DETR models. In this paper, we present Lite DETR, a simple yet efficient end-to-end object detection framework that can effectively reduce the GFLOPs of the detection head by 60% while keeping 99% of the original performance. Specifically, we design an efficient encoder block to update high-level features (corresponding to small-resolution feature maps) and low-level features (corresponding to large-resolution feature maps) in an interleaved way. In addition, to better fuse cross-scale features, we develop a key-aware deformable attention to predict more reliable attention weights. Comprehensive experiments validate the effectiveness and efficiency of the proposed Lite DETR, and the efficient encoder strategy can generalize well across existing DETR-based models. The code will be released after the blind review.
****
HelixSurf: A Robust and Efficient Neural Implicit Surface Learning of Indoor Scenes With Iterative Intertwined Regularization
Zhihao Liang, Zhangjin Huang, Changxing Ding, Kui Jia; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13165-13174
Recovery of an underlying scene geometry from multi-view images stands as a long-time challenge in computer vision research. The recent promise leverages neural implicit surface learning and differentiable volume rendering, and achieves both the recovery of scene geometry and synthesis of novel views, where deep priors of neural models are used as an inductive smoothness bias. While promising for object-level surfaces, these methods suffer when coping with complex scene surfaces. In the meanwhile, traditional multi-view stereo can recover the geometry of scenes with rich textures, by globally optimizing the local, pixel-wise correspondences across multiple views. We are thus motivated to make use of the complementary benefits from the two strategies, and propose a method termed Helix-shaped neural implicit Surface learning or HelixSurf; HelixSurf uses the intermediate prediction from one strategy as the guidance to regularize the learning of the other one, and conducts such intertwined regularization iteratively during the learning process. We also propose an efficient scheme for differentiable volume rendering in HelixSurf. Experiments on surface reconstruction of indoor scenes show that our method compares favorably with existing methods and is orders of magnitude faster, even when some of existing methods are assisted with auxiliary training data. The source code is available at https://github.com/Gorilla-Lab-SCUT/HelixSurf.
****
Joint Appearance and Motion Learning for Efficient Rolling Shutter Correction
Bin Fan, Yuxin Mao, Yuchao Dai, Zhexiong Wan, Qi Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5671-5681
Rolling shutter correction (RSC) is becoming increasingly popular for RS cameras that are widely used in commercial and industrial applications. Despite the promising performance, existing RSC methods typically employ a two-stage network structure that ignores intrinsic information interactions and hinders fast inference. In this paper, we propose a single-stage encoder-decoder-based network, named JAMNet, for efficient RSC. It first extracts pyramid features from consecutive RS inputs, and then simultaneously refines the two complementary information (i.e., global shutter appearance and undistortion motion field) to achieve mutual promotion in a joint learning decoder. To inject sufficient motion cues for guiding joint learning, we introduce a transformer-based motion embedding module and

propose to pass hidden states across pyramid levels. Moreover, we present a new data augmentation strategy "vertical flip + inverse order" to release the potential of the RSC datasets. Experiments on various benchmarks show that our approach surpasses the state-of-the-art methods by a large margin, especially with a 4.7 dB PSNR leap on real-world RSC. Code is available at https://github.com/GitCVfb/JAMNet.
****
Towards a Smaller Student: Capacity Dynamic Distillation for Efficient Image Retrieval
Yi Xie, Huaidong Zhang, Xuemiao Xu, Jianqing Zhu, Shengfeng He; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16006-16015
Previous Knowledge Distillation based efficient image retrieval methods employ a lightweight network as the student model for fast inference. However, the lightweight student model lacks adequate representation capacity for effective knowledge imitation during the most critical early training period, causing final performance degeneration. To tackle this issue, we propose a Capacity Dynamic Distillation framework, which constructs a student model with editable representation capacity. Specifically, the employed student model is initially a heavy model to fruitfully learn distilled knowledge in the early training epochs, and the student model is gradually compressed during the training. To dynamically adjust the model capacity, our dynamic framework inserts a learnable convolutional layer within each residual block in the student model as the channel importance indicator. The indicator is optimized simultaneously by the image retrieval loss and the compression loss, and a retrieval-guided gradient resetting mechanism is proposed to release the gradient conflict. Extensive experiments show that our method has superior inference speed and accuracy, e.g., on the VeRi-776 dataset, given the ResNet101 as a teacher, our method saves 67.13% model parameters and 65.67% FLOPs without sacrificing accuracy.
****
Federated Incremental Semantic Segmentation
Jiahua Dong, Duzhen Zhang, Yang Cong, Wei Cong, Henghui Ding, Dengxin Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3934-3943
Federated learning-based semantic segmentation (FSS) has drawn widespread attention via decentralized training on local clients. However, most FSS models assume categories are fxed in advance, thus heavily undergoing forgetting on old categories in practical applications where local clients receive new categories incrementally while have no memory storage to access old classes. Moreover, new clients collecting novel classes may join in the global training of FSS, which further exacerbates catastrophic forgetting. To surmount the above challenges, we propose a Forgetting-Balanced Learning (FBL) model to address heterogeneous forgetting on old classes from both intra-client and inter-client aspects. Specifically, under the guidance of pseudo labels generated via adaptive class-balanced pseudo labeling, we develop a forgetting-balanced semantic compensation loss and a forgetting-balanced relation consistency loss to rectify intra-client heterogeneous forgetting of old categories with background shift. It performs balanced gradient propagation and relation consistency distillation within local clients. Moreover, to tackle heterogeneous forgetting from inter-client aspect, we propose a task transition monitor. It can identify new classes under privacy protection and store the latest old global model for relation distillation. Qualitative experiments reveal large improvement of our model against comparison methods. The code is available at https://github.com/JiahuaDong/FISS.
****
3D-Aware Facial Landmark Detection via Multi-View Consistent Training on Synthetic Data
Libing Zeng, Lele Chen, Wentao Bao, Zhong Li, Yi Xu, Junsong Yuan, Nima Khademi Kalantari; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12747-12758
Accurate facial landmark detection on wild images plays an essential role in hum

an-computer interaction, entertainment, and medical applications. Existing approaches have limitations in enforcing 3D consistency while detecting 3D/2D facial landmarks due to the lack of multi-view in-the-wild training data. Fortunately, with the recent advances in generative visual models and neural rendering, we have witnessed rapid progress towards high quality 3D image synthesis. In this work, we leverage such approaches to construct a synthetic dataset and propose a novel multi-view consistent learning strategy to improve 3D facial landmark detection accuracy on in-the-wild images. The proposed 3D-aware module can be plugged into any learning-based landmark detection algorithm to enhance its accuracy. We demonstrate the superiority of the proposed plug-in module with extensive comparison against state-of-the-art methods on several real and synthetic datasets.
****
Attention-Based Point Cloud Edge Sampling
Chengzhi Wu, Junwei Zheng, Julius Pfrommer, Jürgen Beyerer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5333-5343
Point cloud sampling is a less explored research topic for this data representation. The most commonly used sampling methods are still classical random sampling and farthest point sampling. With the development of neural networks, various methods have been proposed to sample point clouds in a task-based learning manner. However, these methods are mostly generative-based, rather than selecting points directly using mathematical statistics. Inspired by the Canny edge detection algorithm for images and with the help of the attention mechanism, this paper proposes a non-generative Attention-based Point cloud Edge Sampling method (APES), which captures salient points in the point cloud outline. Both qualitative and quantitative experimental results show the superior performance of our sampling method on common benchmark tasks.
****
Avatars Grow Legs: Generating Smooth Human Motion From Sparse Tracking Inputs With Diffusion Model
Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, Artsiom Sanakoyeu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 481-490
With the recent surge in popularity of AR/VR applications, realistic and accurate control of 3D full-body avatars has become a highly demanded feature. A particular challenge is that only a sparse tracking signal is available from standalone HMDs (Head Mounted Devices), often limited to tracking the user's head and wrists. While this signal is resourceful for reconstructing the upper body motion, the lower body is not tracked and must be synthesized from the limited information provided by the upper body joints. In this paper, we present AGRoL, a novel conditional diffusion model specifically designed to track full bodies given sparse upper-body tracking signals. Our model is based on a simple multi-layer perceptron (MLP) architecture and a novel conditioning scheme for motion data. It can predict accurate and smooth full-body motion, particularly the challenging lower body movement. Unlike common diffusion architectures, our compact architecture can run in real-time, making it suitable for online body-tracking applications. We train and evaluate our model on AMASS motion capture dataset, and demonstrate that our approach outperforms state-of-the-art methods in generated motion accuracy and smoothness. We further justify our design choices through extensive experiments and ablation studies.
****
MobileNeRF: Exploiting the Polygon Rasterization Pipeline for Efficient Neural Field Rendering on Mobile Architectures
Zhiqin Chen, Thomas Funkhouser, Peter Hedman, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16569-16578
Neural Radiance Fields (NeRFs) have demonstrated amazing ability to synthesize images of 3D scenes from novel views. However, they rely upon specialized volumetric rendering algorithms based on ray marching that are mismatched to the capabilities of widely deployed graphics hardware. This paper introduces a new NeRF re

presentation based on textured polygons that can synthesize novel images efficiently with standard rendering pipelines. The NeRF is represented as a set of polygons with textures representing binary opacities and feature vectors. Traditional rendering of the polygons with a z-buffer yields an image with features at every pixel, which are interpreted by a small, view-dependent MLP running in a fragment shader to produce a final pixel color. This approach enables NeRFs to be rendered with the traditional polygon rasterization pipeline, which provides massive pixel-level parallelism, achieving interactive frame rates on a wide range of compute platforms, including mobile phones.
****
Pseudo-Label Guided Contrastive Learning for Semi-Supervised Medical Image Segmentation
Hritam Basak, Zhaozheng Yin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19786-19797
Although recent works in semi-supervised learning (SemiSL) have accomplished significant success in natural image segmentation, the task of learning discriminative representations from limited annotations has been an open problem in medical images. Contrastive Learning (CL) frameworks use the notion of similarity measure which is useful for classification problems, however, they fail to transfer these quality representations for accurate pixel-level segmentation. To this end, we propose a novel semi-supervised patch-based CL framework for medical image segmentation without using any explicit pretext task. We harness the power of both CL and SemiSL, where the pseudo-labels generated from SemiSL aid CL by providing additional guidance, whereas discriminative class information learned in CL leads to accurate multi-class segmentation. Additionally, we formulate a novel loss that synergistically encourages inter-class separability and intra-class compactness among the learned representations. A new inter-patch semantic disparity mapping using average patch entropy is employed for a guided sampling of positives and negatives in the proposed CL framework. Experimental analysis on three publicly available datasets of multiple modalities reveals the superiority of our proposed method as compared to the state-of-the-art methods. Code is available at: https://github.com/hritam-98/PatchCL-MedSeg.
****
Learning Neural Proto-Face Field for Disentangled 3D Face Modeling in the Wild
Zhenyu Zhang, Renwang Chen, Weijian Cao, Ying Tai, Chengjie Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 382-393
Generative models show good potential for recovering 3D faces beyond limited shape assumptions. While plausible details and resolutions are achieved, these models easily fail under extreme conditions of pose, shadow or appearance, due to the entangled fitting or lack of multi-view priors. To address this problem, this paper presents a novel Neural Proto-face Field (NPF) for unsupervised robust 3D face modeling. Instead of using constrained images as Neural Radiance Field (NeRF), NPF disentangles the common/specific facial cues, i.e., ID, expression and scene-specific details from in-the-wild photo collections. Specifically, NPF learns a face prototype to aggregate 3D-consistent identity via uncertainty modeling, extracting multi-image priors from a photo collection. NPF then learns to deform the prototype with the appropriate facial expressions, constrained by a loss of expression consistency and personal idiosyncrasies. Finally, NPF is optimized to fit a target image in the collection, recovering specific details of appearance and geometry. In this way, the generative model benefits from multi-image priors and meaningful facial structures. Extensive experiments on benchmarks show that NPF recovers superior or competitive facial shapes and textures, compared to state-of-the-art methods.
****
Self-Supervised Geometry-Aware Encoder for Style-Based 3D GAN Inversion
Yushi Lan, Xuyi Meng, Shuai Yang, Chen Change Loy, Bo Dai; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20940-20949
StyleGAN has achieved great progress in 2D face reconstruction and semantic edit

ing via image inversion and latent editing. While studies over extending 2D StyleGAN to 3D faces have emerged, a corresponding generic 3D GAN inversion framework is still missing, limiting the applications of 3D face reconstruction and semantic editing. In this paper, we study the challenging problem of 3D GAN inversion where a latent code is predicted given a single face image to faithfully recover its 3D shapes and detailed textures. The problem is ill-posed: innumerable compositions of shape and texture could be rendered to the current image. Furthermore, with the limited capacity of a global latent code, 2D inversion methods cannot preserve faithful shape and texture at the same time when applied to 3D models. To solve this problem, we devise an effective self-training scheme to constrain the learning of inversion. The learning is done efficiently without any real-world 2D-3D training pairs but proxy samples generated from a 3D GAN. In addition, apart from a global latent code that captures the coarse shape and texture information, we augment the generation network with a local branch, where pixel-aligned features are added to faithfully reconstruct face details. We further consider a new pipeline to perform 3D view-consistent editing. Extensive experiments show that our method outperforms state-of-the-art inversion methods in both shape and texture reconstruction quality.
****
PC2: Projection-Conditioned Point Cloud Diffusion for Single-Image 3D Reconstruction
Luke Melas-Kyriazi, Christian Rupprecht, Andrea Vedaldi; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12923-12932
Reconstructing the 3D shape of an object from a single RGB image is a long-standing problem in computer vision. In this paper, we propose a novel method for single-image 3D reconstruction which generates a sparse point cloud via a conditional denoising diffusion process. Our method takes as input a single RGB image along with its camera pose and gradually denoises a set of 3D points, whose positions are initially sampled randomly from a three-dimensional Gaussian distribution, into the shape of an object. The key to our method is a geometrically-consistent conditioning process which we call projection conditioning: at each step in the diffusion process, we project local image features onto the partially-denoised point cloud from the given camera pose. This projection conditioning process enables us to generate high-resolution sparse geometries that are well-aligned with the input image and can additionally be used to predict point colors after shape reconstruction. Moreover, due to the probabilistic nature of the diffusion process, our method is naturally capable of generating multiple different shapes consistent with a single input image. In contrast to prior work, our approach not only performs well on synthetic benchmarks but also gives large qualitative improvements on complex real-world data.
****
Gradient-Based Uncertainty Attribution for Explainable Bayesian Deep Learning
Hanjing Wang, Dhiraj Joshi, Shiqiang Wang, Qiang Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12044-12053
Predictions made by deep learning models are prone to data perturbations, adversarial attacks, and out-of-distribution inputs. To build a trusted AI system, it is therefore critical to accurately quantify the prediction uncertainties. While current efforts focus on improving uncertainty quantification accuracy and efficiency, there is a need to identify uncertainty sources and take actions to mitigate their effects on predictions. Therefore, we propose to develop explainable and actionable Bayesian deep learning methods to not only perform accurate uncertainty quantification but also explain the uncertainties, identify their sources, and propose strategies to mitigate the uncertainty impacts. Specifically, we introduce a gradient-based uncertainty attribution method to identify the most problematic regions of the input that contribute to the prediction uncertainty. Compared to existing methods, the proposed UA-Backprop has competitive accuracy, relaxed assumptions, and high efficiency. Moreover, we propose an uncertainty mitigation strategy that leverages the attribution results as attention to further

improve the model performance. Both qualitative and quantitative evaluations are conducted to demonstrate the effectiveness of our proposed methods.
****

Manipulating Transfer Learning for Property Inference
Yulong Tian, Fnu Suya, Anshuman Suri, Fengyuan Xu, David Evans; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15975-15984
Transfer learning is a popular method for tuning pretrained (upstream) models for different downstream tasks using limited data and computational resources. We study how an adversary with control over an upstream model used in transfer learning can conduct property inference attacks on a victim's tuned downstream model. For example, to infer the presence of images of a specific individual in the downstream training set. We demonstrate attacks in which an adversary can manipulate the upstream model to conduct highly effective and specific property inference attacks (AUC score > 0.9), without incurring significant performance loss on the main task. The main idea of the manipulation is to make the upstream model generate activations (intermediate features) with different distributions for samples with and without a target property, thus enabling the adversary to distinguish easily between downstream models trained with and without training examples that have the target property. Our code is available at https://github.com/yulongt23/Transfer-Inference.
****

POEM: Reconstructing Hand in a Point Embedded Multi-View Stereo
Lixin Yang, Jian Xu, Licheng Zhong, Xinyu Zhan, Zhicheng Wang, Kejian Wu, Cewu Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21108-21117
Enable neural networks to capture 3D geometrical-aware features is essential in multi-view based vision tasks. Previous methods usually encode the 3D information of multi-view stereo into the 2D features. In contrast, we present a novel method, named POEM, that directly operates on the 3D POints Embedded in the Multi-view stereo for reconstructing hand mesh in it. Point is a natural form of 3D information and an ideal medium for fusing features across views, as it has different projections on different views. Our method is thus in light of a simple yet effective idea, that a complex 3D hand mesh can be represented by a set of 3D points that 1) are embedded in the multi-view stereo, 2) carry features from the multi-view images, and 3) encircle the hand. To leverage the power of points, we design two operations: point-based feature fusion and cross-set point attention mechanism. Evaluation on three challenging multi-view datasets shows that POEM outperforms the state-of-the-art in hand mesh reconstruction. Code and models are available for research at github.com/lixiny/POEM
****

BUFFER: Balancing Accuracy, Efficiency, and Generalizability in Point Cloud Registration
Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, Yulan Guo; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1255-1264
An ideal point cloud registration framework should have superior accuracy, acceptable efficiency, and strong generalizability. However, this is highly challenging since existing registration techniques are either not accurate enough, far from efficient, or generalized poorly. It remains an open question that how to achieve a satisfying balance between this three key elements. In this paper, we propose BUFFER, a point cloud registration method for balancing accuracy, efficiency, and generalizability. The key to our approach is to take advantage of both point-wise and patch-wise techniques, while overcoming the inherent drawbacks simultaneously. Different from a simple combination of existing methods, each component of our network has been carefully crafted to tackle specific issues. Specifically, a Point-wise Learner is first introduced to enhance computational efficiency by predicting keypoints and improving the representation capacity of features by estimating point orientations, a Patch-wise Embedder which leverages a lightweight local feature learner is then deployed to extract efficient and general

patch features. Additionally, an Inliers Generator which combines simple neural layers and general features is presented to search inlier correspondences. Exten sive experiments on real-world scenarios demonstrate that our method achieves th e best of both worlds in accuracy, efficiency, and generalization. In particular , our method not only reaches the highest success rate on unseen domains, but al so is almost 30 times faster than the strong baselines specializing in generaliz ation. Code is available at https://github.com/aosheng1996/BUFFER.
****
CrOC: Cross-View Online Clustering for Dense Visual Representation Learning
Thomas Stegmüller, Tim Lebailly, Behzad Bozorgtabar, Tinne Tuytelaars, Jean-Phil ippe Thiran; Proceedings of the IEEE/CVF Conference on Computer Vision and Patte rn Recognition (CVPR), 2023, pp. 7000-7009
Learning dense visual representations without labels is an arduous task and more so from scene-centric data. We propose to tackle this challenging problem by pr oposing a Cross-view consistency objective with an Online Clustering mechanism (CrOC) to discover and segment the semantics of the views. In the absence of hand -crafted priors, the resulting method is more generalizable and does not require a cumbersome pre-processing step. More importantly, the clustering algorithm co njointly operates on the features of both views, thereby elegantly bypassing the issue of content not represented in both views and the ambiguous matching of ob jects from one crop to the other. We demonstrate excellent performance on linear and unsupervised segmentation transfer tasks on various datasets and similarly for video object segmentation. Our code and pre-trained models are publicly avai lable at https://github.com/stegmuel/CrOC.
****
Class Adaptive Network Calibration
Bingyuan Liu, Jérôme Rony, Adrian Galdran, Jose Dolz, Ismail Ben Ayed; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16070-16079
Recent studies have revealed that, beyond conventional accuracy, calibration sho uld also be considered for training modern deep neural networks. To address misc alibration during learning, some methods have explored different penalty functio ns as part of the learning objective, alongside a standard classification loss, with a hyper-parameter controlling the relative contribution of each term. Never theless, these methods share two major drawbacks: 1) the scalar balancing weight is the same for all classes, hindering the ability to address different intrins ic difficulties or imbalance among classes; and 2) the balancing weight is usual ly fixed without an adaptive strategy, which may prevent from reaching the best compromise between accuracy and calibration, and requires hyper-parameter search for each application. We propose Class Adaptive Label Smoothing (CALS) for cali brating deep networks, which allows to learn class-wise multipliers during train ing, yielding a powerful alternative to common label smoothing penalties. Our me thod builds on a general Augmented Lagrangian approach, a well-established techn ique in constrained optimization, but we introduce several modifications to tail or it for large-scale, class-adaptive training. Comprehensive evaluation and mul tiple comparisons on a variety of benchmarks, including standard and long-tailed image classification, semantic segmentation, and text classification, demonstra te the superiority of the proposed method. The code is available at https://gith ub.com/by-liu/CALS.
****
DrapeNet: Garment Generation and Self-Supervised Draping
Luca De Luigi, Ren Li, Benoît Guillard, Mathieu Salzmann, Pascal Fua; Proceeding s of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1451-1460
Recent approaches to drape garments quickly over arbitrary human bodies leverage self-supervision to eliminate the need for large training sets. However, they a re designed to train one network per clothing item, which severely limits their generalization abilities. In our work, we rely on self-supervision to train a si ngle network to drape multiple garments. This is achieved by predicting a 3D def ormation field conditioned on the latent codes of a generative network, which mo

dels garments as unsigned distance fields. Our pipeline can generate and drape p reviously unseen garments of any topology, whose shape can be edited by manipula ting their latent codes. Being fully differentiable, our formulation makes it po ssible to recover accurate 3D models of garments from partial observations -- im ages or 3D scans -- via gradient descent. Our code is publicly available at http s://github.com/liren2515/DrapeNet.
****
Evading Forensic Classifiers With Attribute-Conditioned Adversarial Faces
Fahad Shamshad, Koushik Srivatsan, Karthik Nandakumar; Proceedings of the IEEE/C VF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16469 -16478
The ability of generative models to produce highly realistic synthetic face imag es has raised security and ethical concerns. As a first line of defense against such fake faces, deep learning based forensic classifiers have been developed. W hile these forensic models can detect whether a face image is synthetic or real with high accuracy, they are also vulnerable to adversarial attacks. Although su ch attacks can be highly successful in evading detection by forensic classifiers , they introduce visible noise patterns that are detectable through careful huma n scrutiny. Additionally, these attacks assume access to the target model(s) whi ch may not always be true. Attempts have been made to directly perturb the laten t space of GANs to produce adversarial fake faces that can circumvent forensic c lassifiers. In this work, we go one step further and show that it is possible to successfully generate adversarial fake faces with a specified set of attributes (e.g., hair color, eye size, race, gender, etc.). To achieve this goal, we leve rage the state-of-the-art generative model StyleGAN with disentangled representa tions, which enables a range of modifications without leaving the manifold of na tural images. We propose a framework to search for adversarial latent codes with in the feature space of StyleGAN, where the search can be guided either by a tex t prompt or a reference image. We also propose a meta-learning based optimizatio n strategy to achieve transferable performance on unknown target models. Extensi ve experiments demonstrate that the proposed approach can produce semantically m anipulated adversarial fake faces, which are true to the specified attribute set and can successfully fool forensic face classifiers, while remaining undetectab le by humans. Code: https://github.com/koushiksrivats/face_attribute_attack.
****
FeatureBooster: Boosting Feature Descriptors With a Lightweight Neural Network
Xinjiang Wang, Zeyu Liu, Yu Hu, Wei Xi, Wenxian Yu, Danping Zou; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7630-7639
We introduce a lightweight network to improve descriptors of keypoints within th e same image. The network takes the original descriptors and the geometric prope rties of keypoints as the input, and uses an MLP-based self-boosting stage and a Transformer-based cross-boosting stage to enhance the descriptors. The boosted descriptors can be either real-valued or binary ones. We use the proposed networ k to boost both hand-crafted (ORB, SIFT) and the state-of-the-art learning-based descriptors (SuperPoint, ALIKE) and evaluate them on image matching, visual loc alization, and structure-from-motion tasks. The results show that our method sig nificantly improves the performance of each task, particularly in challenging ca ses such as large illumination changes or repetitive patterns. Our method requir es only 3.2ms on desktop GPU and 27ms on embedded GPU to process 2000 features, which is fast enough to be applied to a practical system. The code and trained w eights are publicly available at github.com/SJTU-ViSYS/FeatureBooster.
****
Progressively Optimized Local Radiance Fields for Robust View Synthesis
Andréas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, Johannes Kopf; Proceedings of the IEEE/CVF Conference on Computer Vision and Pat tern Recognition (CVPR), 2023, pp. 16539-16548
We present an algorithm for reconstructing the radiance field of a large-scale s cene from a single casually captured video. The task poses two core challenges. First, most existing radiance field reconstruction approaches rely on accurate p

re-estimated camera poses from Structure-from-Motion algorithms, which frequentl
y fail on in-the-wild videos. Second, using a single, global radiance field with
 finite representational capacity does not scale to longer trajectories in an un
bounded scene. For handling unknown poses, we jointly estimate the camera poses
with radiance field in a progressive manner. We show that progressive optimizati
on significantly improves the robustness of the reconstruction. For handling lar
ge unbounded scenes, we dynamically allocate new local radiance fields trained w
ith frames within a temporal window. This further improves robustness (e.g., per
forms well even under moderate pose drifts) and allows us to scale to large scen
es. Our extensive evaluation on the Tanks and Temples dataset and our collected
outdoor dataset, Static Hikes, show that our approach compares favorably with th
e state-of-the-art.
****
Towards Efficient Use of Multi-Scale Features in Transformer-Based Object Detect
ors
Gongjie Zhang, Zhipeng Luo, Zichen Tian, Jingyi Zhang, Xiaoqin Zhang, Shijian Lu
; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognit
ion (CVPR), 2023, pp. 6206-6216
Multi-scale features have been proven highly effective for object detection but
often come with huge and even prohibitive extra computation costs, especially fo
r the recent Transformer-based detectors. In this paper, we propose Iterative Mu
lti-scale Feature Aggregation (IMFA) - a generic paradigm that enables efficient
 use of multi-scale features in Transformer-based object detectors. The core ide
a is to exploit sparse multi-scale features from just a few crucial locations, a
nd it is achieved with two novel designs. First, IMFA rearranges the Transformer
 encoder-decoder pipeline so that the encoded features can be iteratively update
d based on the detection predictions. Second, IMFA sparsely samples scale-adapti
ve features for refined detection from just a few keypoint locations under the g
uidance of prior detection predictions. As a result, the sampled multi-scale fea
tures are sparse yet still highly beneficial for object detection. Extensive exp
eriments show that the proposed IMFA boosts the performance of multiple Transfor
mer-based object detectors significantly yet with only slight computational over
head.
****
Delivering Arbitrary-Modal Semantic Segmentation
Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodon
g Fu, Kaiwei Wang, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Conference o
n Computer Vision and Pattern Recognition (CVPR), 2023, pp. 1136-1147
Multimodal fusion can make semantic segmentation more robust. However, fusing an
 arbitrary number of modalities remains underexplored. To delve into this proble
m, we create the DeLiVER arbitrary-modal segmentation benchmark, covering Depth,
 LiDAR, multiple Views, Events, and RGB. Aside from this, we provide this datase
t in four severe weather conditions as well as five sensor failure cases to expl
oit modal complementarity and resolve partial outages. To facilitate this data,
we present the arbitrary cross-modal segmentation model CMNeXt. It encompasses a
 Self-Query Hub (SQ-Hub) designed to extract effective information from any moda
lity for subsequent fusion with the RGB representation and adds only negligible
amounts of parameters (0.01M) per additional modality. On top, to efficiently a
nd flexibly harvest discriminative cues from the auxiliary modalities, we introd
uce the simple Parallel Pooling Mixer (PPX). With extensive experiments on a tot
al of six benchmarks, our CMNeXt achieves state-of-the-art performance, allowing
 to scale from 1 to 81 modalities on the DeLiVER, KITTI-360, MFNet, NYU Depth V2
, UrbanLF, and MCubeS datasets. On the freshly collected DeLiVER, the quad-modal
 CMNeXt reaches up to 66.30% in mIoU with a +9.10% gain as compared to the mono-
modal baseline.
****
GeoMVSNet: Learning Multi-View Stereo With Geometry Perception
Zhe Zhang, Rui Peng, Yuxi Hu, Ronggang Wang; Proceedings of the IEEE/CVF Confere
nce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21508-21518
Recent cascade Multi-View Stereo (MVS) methods can efficiently estimate high-res

olution depth maps through narrowing hypothesis ranges. However, previous method s ignored the vital geometric information embedded in coarse stages, leading to vulnerable cost matching and sub-optimal reconstruction results. In this paper, we propose a geometry awareness model, termed GeoMVSNet, to explicitly integrate geometric clues implied in coarse stages for delicate depth estimation. In particular, we design a two-branch geometry fusion network to extract geometric priors from coarse estimations to enhance structural feature extraction at finer stages. Besides, we embed the coarse probability volumes, which encode valuable depth distribution attributes, into the lightweight regularization network to further strengthen depth-wise geometry intuition. Meanwhile, we apply the frequency domain filtering to mitigate the negative impact of the high-frequency regions and adopt the curriculum learning strategy to progressively boost the geometry integration of the model. To intensify the full-scene geometry perception of our model, we present the depth distribution similarity loss based on the Gaussian-Mixture Model assumption. Extensive experiments on DTU and Tanks and Temples (T&T) datasets demonstrate that our GeoMVSNet achieves state-of-the-art results and ranks first on the T&T-Advanced set. Code is available at https://github.com/doublez0108/GeoMVSNet.
****
Consistent-Teacher: Towards Reducing Inconsistent Pseudo-Targets in Semi-Supervised Object Detection
Xinjiang Wang, Xingyi Yang, Shilong Zhang, Yijiang Li, Litong Feng, Shijie Fang, Chengqi Lyu, Kai Chen, Wayne Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3240-3249
In this study, we dive deep into the inconsistency of pseudo targets in semi-supervised object detection (SSOD). Our core observation is that the oscillating pseudo-targets undermine the training of an accurate detector. It injects noise into the student's training, leading to severe overfitting problems. Therefore, we propose a systematic solution, termed NAME, to reduce the inconsistency. First, adaptive anchor assignment (ASA) substitutes the static IoU-based strategy, which enables the student network to be resistant to noisy pseudo-bounding boxes. Then we calibrate the subtask predictions by designing a 3D feature alignment module (FAM-3D). It allows each classification feature to adaptively query the optimal feature vector for the regression task at arbitrary scales and locations. Lastly, a Gaussian Mixture Model (GMM) dynamically revises the score threshold of pseudo-bboxes, which stabilizes the number of ground truths at an early stage and remedies the unreliable supervision signal during training. NAME provides strong results on a large range of SSOD evaluations. It achieves 40.0 mAP with ResNet-50 backbone given only 10% of annotated MS-COCO data, which surpasses previous baselines using pseudo labels by around 3 mAP. When trained on fully annotated MS-COCO with additional unlabeled data, the performance further increases to 47.7 mAP. Our code is available at https://github.com/Adamdad/ConsistentTeacher.
****
OCTET: Object-Aware Counterfactual Explanations
Mehdi Zemni, Mickaël Chen, Éloi Zablocki, Hédi Ben-Younes, Patrick Pérez, Matthieu Cord; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15062-15071
Nowadays, deep vision models are being widely deployed in safety-critical applications, e.g., autonomous driving, and explainability of such models is becoming a pressing concern. Among explanation methods, counterfactual explanations aim to find minimal and interpretable changes to the input image that would also change the output of the model to be explained. Such explanations point end-users at the main factors that impact the decision of the model. However, previous methods struggle to explain decision models trained on images with many objects, e.g., urban scenes, which are more difficult to work with but also arguably more critical to explain. In this work, we propose to tackle this issue with an object-centric framework for counterfactual explanation generation. Our method, inspired by recent generative modeling works, encodes the query image into a latent space that is structured in a way to ease object-level manipulations. Doing so, it provides the end-user with control over which search directions (e.g., spatial di

splacement of objects, style modification, etc.) are to be explored during the c
ounterfactual generation. We conduct a set of experiments on counterfactual expl
anation benchmarks for driving scenes, and we show that our method can be adapte
d beyond classification, e.g., to explain semantic segmentation models. To compl
ete our analysis, we design and run a user study that measures the usefulness of
 counterfactual explanations in understanding a decision model. Code is availabl
e at https://github.com/valeoai/OCTET.
****
TeSLA: Test-Time Self-Learning With Automatic Adversarial Augmentation
Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, Jean-Philippe Thiran; Procee
dings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVP
R), 2023, pp. 20341-20350
Most recent test-time adaptation methods focus on only classification tasks, use
 specialized network architectures, destroy model calibration or rely on lightwe
ight information from the source domain. To tackle these issues, this paper prop
oses a novel Test-time Self-Learning method with automatic Adversarial augmentat
ion dubbed TeSLA for adapting a pre-trained source model to the unlabeled stream
ing test data. In contrast to conventional self-learning methods based on cross-
entropy, we introduce a new test-time loss function through an implicitly tight
connection with the mutual information and online knowledge distillation. Furthe
rmore, we propose a learnable efficient adversarial augmentation module that fur
ther enhances online knowledge distillation by simulating high entropy augmented
 images. Our method achieves state-of-the-art classification and segmentation re
sults on several benchmarks and types of domain shifts, particularly on challeng
ing measurement shifts of medical images. TeSLA also benefits from several desir
able properties compared to competing methods in terms of calibration, uncertain
ty metrics, insensitivity to model architectures, and source training strategies
, all supported by extensive ablations. Our code and models are available at htt
ps://github.com/devavratTomar/TeSLA.
****
DNeRV: Modeling Inherent Dynamics via Difference Neural Representation for Video
s
Qi Zhao, M. Salman Asif, Zhan Ma; Proceedings of the IEEE/CVF Conference on Comp
uter Vision and Pattern Recognition (CVPR), 2023, pp. 2031-2040
Existing implicit neural representation (INR) methods do not fully exploit spati
otemporal redundancies in videos. Index-based INRs ignore the content-specific s
patial features and hybrid INRs ignore the contextual dependency on adjacent fra
mes, leading to poor modeling capability for scenes with large motion or dynamic
s. We analyze this limitation from the perspective of function fitting and revea
l the importance of frame difference. To use explicit motion information, we pro
pose Difference Neural Representation for Videos (DNeRV), which consists of two
streams for content and frame difference. We also introduce a collaborative cont
ent unit for effective feature fusion. We test DNeRV for video compression, inpa
inting, and interpolation. DNeRV achieves competitive results against the state-
of-the-art neural compression approaches and outperforms existing implicit metho
ds on downstream inpainting and interpolation for 960 x 1920 videos.
****
RefTeacher: A Strong Baseline for Semi-Supervised Referring Expression Comprehen
sion
Jiamu Sun, Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, Rongron
g Ji; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Reco
gnition (CVPR), 2023, pp. 19144-19154
Referring expression comprehension (REC) often requires a large number of instan
ce-level annotations for fully supervised learning, which are laborious and expe
nsive. In this paper, we present the first attempt of semi-supervised learning f
or REC and propose a strong baseline method called RefTeacher. Inspired by the r
ecent progress in computer vision, RefTeacher adopts a teacher-student learning
paradigm, where the teacher REC network predicts pseudo-labels for optimizing th
e student one. This paradigm allows REC models to exploit massive unlabeled data
 based on a small fraction of labeled. In particular, we also identify two key c

hallenges in semi-supervised REC, namely, sparse supervision signals and worse p
seudo-label noise. To address these issues, we equip RefTeacher with two novel d
esigns called Attention-based Imitation Learning (AIL) and Adaptive Pseudo-label
 Weighting (APW). AIL can help the student network imitate the recognition behav
iors of the teacher, thereby obtaining sufficient supervision signals. APW can h
elp the model adaptively adjust the contributions of pseudo-labels with varying
qualities, thus avoiding confirmation bias. To validate RefTeacher, we conduct e
xtensive experiments on three REC benchmark datasets. Experimental results show
that RefTeacher obtains obvious gains over the fully supervised methods. More im
portantly, using only 10% labeled data, our approach allows the model to achieve
 near 100% fully supervised performance, e.g., only -2.78% on RefCOCO.
****
Handwritten Text Generation From Visual Archetypes
Vittorio Pippi, Silvia Cascianelli, Rita Cucchiara; Proceedings of the IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22458-22
467
Generating synthetic images of handwritten text in a writer-specific style is a
challenging task, especially in the case of unseen styles and new words, and eve
n more when these latter contain characters that are rarely encountered during t
raining. While emulating a writer's style has been recently addressed by generat
ive models, the generalization towards rare characters has been disregarded. In
this work, we devise a Transformer-based model for Few-Shot styled handwritten t
ext generation and focus on obtaining a robust and informative representation of
 both the text and the style. In particular, we propose a novel representation o
f the textual content as a sequence of dense vectors obtained from images of sym
bols written as standard GNU Unifont glyphs, which can be considered their visua
l archetypes. This strategy is more suitable for generating characters that, des
pite having been seen rarely during training, possibly share visual details with
 the frequently observed ones. As for the style, we obtain a robust representati
on of unseen writers' calligraphy by exploiting specific pre-training on a large
 synthetic dataset. Quantitative and qualitative results demonstrate the effecti
veness of our proposal in generating words in unseen styles and with rare charac
ters more faithfully than existing approaches relying on independent one-hot enc
odings of the characters.
****
Unicode Analogies: An Anti-Objectivist Visual Reasoning Challenge
Steven Spratley, Krista A. Ehinger, Tim Miller; Proceedings of the IEEE/CVF Conf
erence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19082-19091
Analogical reasoning enables agents to extract relevant information from scenes,
 and efficiently navigate them in familiar ways. While progressive-matrix proble
ms (PMPs) are becoming popular for the development and evaluation of analogical
reasoning in computer vision, we argue that the dominant methodology in this are
a struggles to expose the lack of meaningful generalisation in solvers, and rein
forces an objectivist stance on perception -- that objects can only be seen one
way -- which we believe to be counter-productive. In this paper, we introduce th
e Unicode Analogies challenge, consisting of polysemic, character-based PMPs to
benchmark fluid conceptualisation ability in vision systems. Writing systems hav
e evolved characters at multiple levels of abstraction, from iconic through to s
ymbolic representations, producing both visually interrelated yet exceptionally
diverse images when compared to those exhibited by existing PMP datasets. Our fr
amework has been designed to challenge models by presenting tasks much harder to
 complete without robust feature extraction, while remaining largely solvable by
 human participants. We therefore argue that Unicode Analogies elegantly capture
s and tests for a facet of human visual reasoning that is severely lacking in cu
rrent-generation AI.
****
FFF: Fragment-Guided Flexible Fitting for Building Complete Protein Structures
Weijie Chen, Xinyan Wang, Yuhang Wang; Proceedings of the IEEE/CVF Conference on
 Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19776-19785
Cryo-electron microscopy (cryo-EM) is a technique for reconstructing the 3-dimen

sional (3D) structure of biomolecules (especially large protein complexes and molecular assemblies). As the resolution increases to the near-atomic scale, building protein structures de novo from cryo-EM maps becomes possible. Recently, recognition-based de novo building methods have shown the potential to streamline this process. However, it cannot build a complete structure due to the low signal-to-noise ratio (SNR) problem. At the same time, AlphaFold has led to a great breakthrough in predicting protein structures. This has inspired us to combine fragment recognition and structure prediction methods to build a complete structure. In this paper, we propose a new method named FFF that bridges protein structure prediction and protein structure recognition with flexible fitting. First, a multi-level recognition network is used to capture various structural features from the input 3D cryo-EM map. Next, protein structural fragments are generated using pseudo peptide vectors and a protein sequence alignment method based on these extracted features. Finally, a complete structural model is constructed using the predicted protein fragments via flexible fitting. Based on our benchmark tests, FFF outperforms the baseline meth- ods for building complete protein structures.
****
Polarized Color Image Denoising
Zhuoxiao Li, Haiyang Jiang, Mingdeng Cao, Yinqiang Zheng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9873-9882
Single-chip polarized color photography provides both visual textures and object surface information in one snapshot. However, the use of an additional directional polarizing filter array tends to lower photon count and SNR, when compared to conventional color imaging. As a result, such a bilayer structure usually leads to unpleasant noisy images and undermines performance of polarization analysis, especially in low-light conditions. It is a challenge for traditional image processing pipelines owing to the fact that the physical constraints exerted implicitly in the channels are excessively complicated. In this paper, we propose to tackle this issue through a noise modeling method for realistic data synthesis and a powerful network structure inspired by vision Transformer. A real-world polarized color image dataset of paired raw short-exposed noisy images and long-exposed reference images is captured for experimental evaluation, which has demonstrated the effectiveness of our approaches for data synthesis and polarized color image denoising.
****
Continuous Pseudo-Label Rectified Domain Adaptive Semantic Segmentation With Implicit Neural Representations
Rui Gong, Qin Wang, Martin Danelljan, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7225-7235
Unsupervised domain adaptation (UDA) for semantic segmentation aims at improving the model performance on the unlabeled target domain by leveraging a labeled source domain. Existing approaches have achieved impressive progress by utilizing pseudo-labels on the unlabeled target-domain images. Yet the low-quality pseudo-labels, arising from the domain discrepancy, inevitably hinder the adaptation. This calls for effective and accurate approaches to estimating the reliability of the pseudo-labels, in order to rectify them. In this paper, we propose to estimate the rectification values of the predicted pseudo-labels with implicit neural representations. We view the rectification value as a signal defined over the continuous spatial domain. Taking an image coordinate and the nearby deep features as inputs, the rectification value at a given coordinate is predicted as an output. This allows us to achieve high-resolution and detailed rectification values estimation, important for accurate pseudo-label generation at mask boundaries in particular. The rectified pseudo-labels are then leveraged in our rectification-aware mixture model (RMM) to be learned end-to-end and help the adaptation. We demonstrate the effectiveness of our approach on different UDA benchmarks, including synthetic-to-real and day-to-night. Our approach achieves superior results compared to state-of-the-art. The implementation is available at https://githu

b.com/ETHRuiGong/IR2F.
****
Hyperbolic Contrastive Learning for Visual Representations Beyond Objects
Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, David Jacobs; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6840-6849
Although self-/un-supervised methods have led to rapid progress in visual representation learning, these methods generally treat objects and scenes using the same lens. In this paper, we focus on learning representations of objects and scenes that preserve the structure among them. Motivated by the observation that visually similar objects are close in the representation space, we argue that the scenes and objects should instead follow a hierarchical structure based on their compositionality. To exploit such a structure, we propose a contrastive learning framework where a Euclidean loss is used to learn object representations and a hyperbolic loss is used to encourage representations of scenes to lie close to representations of their constituent objects in hyperbolic space. This novel hyperbolic objective encourages the scene-object hypernymy among the representations by optimizing the magnitude of their norms. We show that when pretraining on the COCO and OpenImages datasets, the hyperbolic loss improves the downstream performance of several baselines across multiple datasets and tasks, including image classification, object detection, and semantic segmentation. We also show that the properties of the learned representations allow us to solve various vision tasks that involve the interaction between scenes and objects in a zero-shot fashion.
****
Align Your Latents: High-Resolution Video Synthesis With Latent Diffusion Models
Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, Karsten Kreis; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22563-22575
Latent Diffusion Models (LDMs) enable high-quality image synthesis while avoiding excessive compute demands by training a diffusion model in a compressed lower-dimensional latent space. Here, we apply the LDM paradigm to high-resolution video generation, a particularly resource-intensive task. We first pre-train an LDM on images only; then, we turn the image generator into a video generator by introducing a temporal dimension to the latent space diffusion model and fine-tuning on encoded image sequences, i.e., videos. Similarly, we temporally align diffusion model upsamplers, turning them into temporally consistent video super resolution models. We focus on two relevant real-world applications: Simulation of in-the-wild driving data and creative content creation with text-to-video modeling. In particular, we validate our Video LDM on real driving videos of resolution 512x1024, achieving state-of-the-art performance. Furthermore, our approach can easily leverage off-the-shelf pre-trained image LDMs, as we only need to train a temporal alignment model in that case. Doing so, we turn the publicly available, state-of-the-art text-to-image LDM Stable Diffusion into an efficient and expressive text-to-video model with resolution up to 1280x2048. We show that the temporal layers trained in this way generalize to different fine-tuned text-to-image LDMs. Utilizing this property, we show the first results for personalized text-to-video generation, opening exciting directions for future content creation. Project page: https://nv-tlabs.github.io/VideoLDM/
****
AligNeRF: High-Fidelity Neural Radiance Fields via Alignment-Aware Training
Yifan Jiang, Peter Hedman, Ben Mildenhall, Dejia Xu, Jonathan T. Barron, Zhangyang Wang, Tianfan Xue; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 46-55
Neural Radiance Fields (NeRFs) are a powerful representation for modeling a 3D scene as a continuous function. Though NeRF is able to render complex 3D scenes with view-dependent effects, few efforts have been devoted to exploring its limits in a high-resolution setting. Specifically, existing NeRF-based methods face several limitations when reconstructing high-resolution real scenes, including a very large number of parameters, misaligned input data, and overly smooth detail

s. In this work, we conduct the first pilot study on training NeRF with high-resolution data and propose the corresponding solutions: 1) marrying the multilayer perceptron (MLP) with convolutional layers which can encode more neighborhood information while reducing the total number of parameters; 2) a novel training strategy to address misalignment caused by moving objects or small camera calibration errors; and 3) a high-frequency aware loss. Our approach is nearly free without introducing obvious training/testing costs, while experiments on different datasets demonstrate that it can recover more high-frequency details compared with the current state-of-the-art NeRF models. Project page: https://yifanjiang19.github.io/alignerf.
****
NAR-Former: Neural Architecture Representation Learning Towards Holistic Attributes Prediction
Yun Yi, Haokui Zhang, Wenze Hu, Nannan Wang, Xiaoyu Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7715-7724
With the wide and deep adoption of deep learning models in real applications, there is an increasing need to model and learn the representations of the neural networks themselves. These models can be used to estimate attributes of different neural network architectures such as the accuracy and latency, without running the actual training or inference tasks. In this paper, we propose a neural architecture representation model that can be used to estimate these attributes holistically. Specifically, we first propose a simple and effective tokenizer to encode both the operation and topology information of a neural network into a single sequence. Then, we design a multi-stage fusion transformer to build a compact vector representation from the converted sequence. For efficient model training, we further propose an information flow consistency augmentation and correspondingly design an architecture consistency loss, which brings more benefits with less augmentation samples compared with previous random augmentation strategies. Experiment results on NAS-Bench-101, NAS-Bench-201, DARTS search space and NNLQP show that our proposed framework can be used to predict the aforementioned latency and accuracy attributes of both cell architectures and whole deep neural networks, and achieves promising performance. Code is available at https://github.com/yuny220/NAR-Former.
****
Implicit 3D Human Mesh Recovery Using Consistency With Pose and Shape From Unseen-View
Hanbyel Cho, Yooshin Cho, Jaesung Ahn, Junmo Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21148-21158
From an image of a person, we can easily infer the natural 3D pose and shape of the person even if ambiguity exists. This is because we have a mental model that allows us to imagine a person's appearance at different viewing directions from a given image and utilize the consistency between them for inference. However, existing human mesh recovery methods only consider the direction in which the image was taken due to their structural limitations. Hence, we propose "Implicit 3D Human Mesh Recovery (ImpHMR)" that can implicitly imagine a person in 3D space at the feature-level via Neural Feature Fields. In ImpHMR, feature fields are generated by CNN-based image encoder for a given image. Then, the 2D feature map is volume-rendered from the feature field for a given viewing direction, and the pose and shape parameters are regressed from the feature. To utilize consistency with pose and shape from unseen-view, if there are 3D labels, the model predicts results including the silhouette from an arbitrary direction and makes it equal to the rotated ground-truth. In the case of only 2D labels, we perform self-supervised learning through the constraint that the pose and shape parameters inferred from different directions should be the same. Extensive evaluations show the efficacy of the proposed method.
****
UniDAformer: Unified Domain Adaptive Panoptic Segmentation Transformer via Hierarchical Mask Calibration

Jingyi Zhang, Jiaxing Huang, Xiaoqin Zhang, Shijian Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11227-11237
Domain adaptive panoptic segmentation aims to mitigate data annotation challenge by leveraging off-the-shelf annotated data in one or multiple related source domains. However, existing studies employ two separate networks for instance segmentation and semantic segmentation which lead to excessive network parameters as well as complicated and computationally intensive training and inference processes. We design UniDAformer, a unified domain adaptive panoptic segmentation transformer that is simple but can achieve domain adaptive instance segmentation and semantic segmentation simultaneously within a single network. UniDAformer introduces Hierarchical Mask Calibration (HMC) that rectifies inaccurate predictions at the level of regions, superpixels and pixels via online self-training on the fly. It has three unique features: 1) it enables unified domain adaptive panoptic adaptation; 2) it mitigates false predictions and improves domain adaptive panoptic segmentation effectively; 3) it is end-to-end trainable with a much simpler training and inference pipeline. Extensive experiments over multiple public benchmarks show that UniDAformer achieves superior domain adaptive panoptic segmentation as compared with the state-of-the-art.
****
Non-Contrastive Learning Meets Language-Image Pre-Training
Jinghao Zhou, Li Dong, Zhe Gan, Lijuan Wang, Furu Wei; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 11028-11038
Contrastive language-image pre-training (CLIP) serves as a de-facto standard to align images and texts. Nonetheless, the loose correlation between images and texts of web-crawled data renders the contrastive objective data inefficient and craving for a large training batch size. In this work, we explore the validity of non-contrastive language-image pre-training (nCLIP) and study whether nice properties exhibited in visual self-supervised models can emerge. We empirically observe that the non-contrastive objective nourishes representation learning while sufficiently underperforming under zero-shot recognition. Based on the above study, we further introduce xCLIP, a multi-tasking framework combining CLIP and nCLIP, and show that nCLIP aids CLIP in enhancing feature semantics. The synergy between two objectives lets xCLIP enjoy the best of both worlds: superior performance in both zero-shot transfer and representation learning. Systematic evaluation is conducted spanning a wide variety of downstream tasks including zero-shot classification, out-of-domain classification, retrieval, visual representation learning, and textual representation learning, showcasing a consistent performance gain and validating the effectiveness of xCLIP.
****
Teaching Structured Vision & Language Concepts to Vision & Language Models
Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, Leonid Karlinsky; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2657-2668
Vision and Language (VL) models have demonstrated remarkable zero-shot performance in a variety of tasks. However, some aspects of complex language understanding still remain a challenge. We introduce the collective notion of Structured Vision & Language Concepts (SVLC) which includes object attributes, relations, and states which are present in the text and visible in the image. Recent studies have shown that even the best VL models struggle with SVLC. A possible way of fixing this issue is by collecting dedicated datasets for teaching each SVLC type, yet this might be expensive and time-consuming. Instead, we propose a more elegant data-driven approach for enhancing VL models' understanding of SVLCs that makes more effective use of existing VL pre-training datasets and does not require any additional data. While automatic understanding of image structure still remains largely unsolved, language structure is much better modeled and understood, allowing for its effective utilization in teaching VL models. In this paper, we propose various techniques based on language structure understanding that can be

used to manipulate the textual part of off-the-shelf paired VL datasets. VL models trained with the updated data exhibit a significant improvement of up to 15% in their SVLC understanding with only a mild degradation in their zero-shot capabilities both when training from scratch or fine-tuning a pre-trained model. Our code and pretrained models are available at: https://github.com/SivanDoveh/TSVLC
****
Teleidoscopic Imaging System for Microscale 3D Shape Reconstruction
Ryo Kawahara, Meng-Yu Jennifer Kuo, Shohei Nobuhara; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20813-20822
This paper proposes a practical method of microscale 3D shape capturing by a teleidoscopic imaging system. The main challenge in microscale 3D shape reconstruction is to capture the target from multiple viewpoints with a large enough depth-of-field. Our idea is to employ a teleidoscopic measurement system consisting of three planar mirrors and monocentric lens. The planar mirrors virtually define multiple viewpoints by multiple reflections, and the monocentric lens realizes a high magnification with less blurry and surround view even in closeup imaging. Our contributions include, a structured ray-pixel camera model which handles refractive and reflective projection rays efficiently, analytical evaluations of depth of field of our teleidoscopic imaging system, and a practical calibration algorithm of the teleidoscppic imaging system. Evaluations with real images prove the concept of our measurement system.
****
UV Volumes for Real-Time Rendering of Editable Free-View Human Performance
Yue Chen, Xuan Wang, Xingyu Chen, Qi Zhang, Xiaoyu Li, Yu Guo, Jue Wang, Fei Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16621-16631
Neural volume rendering enables photo-realistic renderings of a human performer in free-view, a critical task in immersive VR/AR applications. But the practice is severely limited by high computational costs in the rendering process. To solve this problem, we propose the UV Volumes, a new approach that can render an editable free-view video of a human performer in real-time. It separates the high-frequency (i.e., non-smooth) human appearance from the 3D volume, and encodes them into 2D neural texture stacks (NTS). The smooth UV volumes allow much smaller and shallower neural networks to obtain densities and texture coordinates in 3D while capturing detailed appearance in 2D NTS. For editability, the mapping between the parameterized human model and the smooth texture coordinates allows us a better generalization on novel poses and shapes. Furthermore, the use of NTS enables interesting applications, e.g., retexturing. Extensive experiments on CMU Panoptic, ZJU Mocap, and H36M datasets show that our model can render 960 x 540 images in 30FPS on average with comparable photo-realism to state-of-the-art methods.
****
NULL-Text Inversion for Editing Real Images Using Guided Diffusion Models
Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, Daniel Cohen-Or; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6038-6047
Recent large-scale text-guided diffusion models provide powerful image generation capabilities. Currently, a massive effort is given to enable the modification of these images using text only as means to offer intuitive and versatile editing tools. To edit a real image using these state-of-the-art tools, one must first invert the image with a meaningful text prompt into the pretrained model's domain. In this paper, we introduce an accurate inversion technique and thus facilitate an intuitive text-based modification of the image. Our proposed inversion consists of two key novel components: (i) Pivotal inversion for diffusion models. While current methods aim at mapping random noise samples to a single input image, we use a single pivotal noise vector for each timestamp and optimize around it. We recognize that a direct DDIM inversion is inadequate on its own, but does provide a rather good anchor for our optimization. (ii) NULL-text optimization,

where we only modify the unconditional textual embedding that is used for classifier-free guidance, rather than the input text embedding. This allows for keeping both the model weights and the conditional embedding intact and hence enables applying prompt-based editing while avoiding the cumbersome tuning of the model's weights. Our Null-text inversion, based on the publicly available Stable Diffusion model, is extensively evaluated on a variety of images and various prompt editing, showing high-fidelity editing of real images.
****
JacobiNeRF: NeRF Shaping With Mutual Information Gradients
Xiaomeng Xu, Yanchao Yang, Kaichun Mo, Boxiao Pan, Li Yi, Leonidas Guibas; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 16498-16507
We propose a method that trains a neural radiance field (NeRF) to encode not only the appearance of the scene but also semantic correlations between scene points, regions, or entities -- aiming to capture their mutual co-variation patterns. In contrast to the traditional first-order photometric reconstruction objective, our method explicitly regularizes the learning dynamics to align the Jacobians of highly-correlated entities, which proves to maximize the mutual information between them under random scene perturbations. By paying attention to this second-order information, we can shape a NeRF to express semantically meaningful synergies when the network weights are changed by a delta along the gradient of a single entity, region, or even a point. To demonstrate the merit of this mutual information modeling, we leverage the coordinated behavior of scene entities that emerges from our shaping to perform label propagation for semantic and instance segmentation. Our experiments show that a JacobiNeRF is more efficient in propagating annotations among 2D pixels and 3D points compared to NeRFs without mutual information shaping, especially in extremely sparse label regimes -- thus reducing annotation burden. The same machinery can further be used for entity selection or scene modifications. Our code is available at https://github.com/xxm19/jacobinerf.
****
Selective Structured State-Spaces for Long-Form Video Understanding
Jue Wang, Wentao Zhu, Pichao Wang, Xiang Yu, Linda Liu, Mohamed Omar, Raffay Hamid; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6387-6397
Effective modeling of complex spatiotemporal dependencies in long-form videos remains an open problem. The recently proposed Structured State-Space Sequence (S4) model with its linear complexity offers a promising direction in this space. However, we demonstrate that treating all image-tokens equally as done by S4 model can adversely affect its efficiency and accuracy. To address this limitation, we present a novel Selective S4 (i.e., S5) model that employs a lightweight mask generator to adaptively select informative image tokens resulting in more efficient and accurate modeling of long-term spatiotemporal dependencies in videos. Unlike previous mask-based token reduction methods used in transformers, our S5 model avoids the dense self-attention calculation by making use of the guidance of the momentum-updated S4 model. This enables our model to efficiently discard less informative tokens and adapt to various long-form video understanding tasks more effectively. However, as is the case for most token reduction methods, the informative image tokens could be dropped incorrectly. To improve the robustness and the temporal horizon of our model, we propose a novel long-short masked contrastive learning (LSMCL) approach that enables our model to predict longer temporal context using shorter input videos. We present extensive comparative results using three challenging long-form video understanding datasets (LVU, COIN and Breakfast), demonstrating that our approach consistently outperforms the previous state-of-the-art S4 model by up to 9.6% accuracy while reducing its memory footprint by 23%.
****
Open-Set Representation Learning Through Combinatorial Embedding
Geeho Kim, Junoh Kang, Bohyung Han; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19744-19753

Visual recognition tasks are often limited to dealing with a small subset of classes simply because the labels for the remaining classes are unavailable. We are interested in identifying novel concepts in a dataset through representation learning based on both labeled and unlabeled examples, and extending the horizon of recognition to both known and novel classes. To address this challenging task, we propose a combinatorial learning approach, which naturally clusters the examples in unseen classes using the compositional knowledge given by multiple supervised meta-classifiers on heterogeneous label spaces. The representations given by the combinatorial embedding are made more robust by unsupervised pairwise relation learning. The proposed algorithm discovers novel concepts via a joint optimization for enhancing the discrimitiveness of unseen classes as well as learning the representations of known classes generalizable to novel ones. Our extensive experiments demonstrate remarkable performance gains by the proposed approach on public datasets for image retrieval and image categorization with novel class discovery.

****

Multi-View Stereo Representation Revist: Region-Aware MVSNet

Yisu Zhang, Jianke Zhu, Lixiang Lin; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17376-17385

Deep learning-based multi-view stereo has emerged as a powerful paradigm for reconstructing the complete geometrically-detailed objects from multi-views. Most of the existing approaches only estimate the pixel-wise depth value by minimizing the gap between the predicted point and the intersection of ray and surface, which usually ignore the surface topology. It is essential to the textureless regions and surface boundary that cannot be properly reconstructed.To address this issue, we suggest to take advantage of point-to-surface distance so that the model is able to perceive a wider range of surfaces. To this end, we predict the distance volume from cost volume to estimate the signed distance of points around the surface. Our proposed RA-MVSNet is patch-awared, since the perception range is enhanced by associating hypothetical planes with a patch of surface. Therefore, it could increase the completion of textureless regions and reduce the outliers at the boundary. Moreover, the mesh topologies with fine details can be generated by the introduced distance volume. Comparing to the conventional deep learning-based multi-view stereo methods, our proposed RA-MVSNet approach obtains more complete reconstruction results by taking advantage of signed distance supervision. The experiments on both the DTU and Tanks & Temples datasets demonstrate that our proposed approach achieves the state-of-the-art results.

****

A Unified HDR Imaging Method With Pixel and Patch Level

Qingsen Yan, Weiye Chen, Song Zhang, Yu Zhu, Jinqiu Sun, Yanning Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22211-22220

Mapping Low Dynamic Range (LDR) images with different exposures to High Dynamic Range (HDR) remains nontrivial and challenging on dynamic scenes due to ghosting caused by object motion or camera jitting. With the success of Deep Neural Networks (DNNs), several DNNs-based methods have been proposed to alleviate ghosting, they cannot generate approving results when motion and saturation occur. To generate visually pleasing HDR images in various cases, we propose a hybrid HDR deghosting network, called HyHDRNet, to learn the complicated relationship between reference and non-reference images. The proposed HyHDRNet consists of a content alignment subnetwork and a Transformer-based fusion subnetwork. Specifically, to effectively avoid ghosting from the source, the content alignment subnetwork uses patch aggregation and ghost attention to integrate similar content from other non-reference images with patch level and suppress undesired components with pixel level. To achieve mutual guidance between patch-level and pixel-level, we leverage a gating module to sufficiently swap useful information both in ghosted and saturated regions. Furthermore, to obtain a high-quality HDR image, the Transformer-based fusion subnetwork uses a Residual Deformable Transformer Block (RDTB) to adaptively merge information for different exposed regions. We examined the proposed method on four widely used public HDR image deghosting datasets. Exp

eriments demonstrate that HyHDRNet outperforms state-of-the-art methods both quantitatively and qualitatively, achieving appealing HDR visualization with unified textures and colors.
****
Motion Information Propagation for Neural Video Compression
Linfeng Qi, Jiahao Li, Bin Li, Houqiang Li, Yan Lu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 6111-6120
In most existing neural video codecs, the information flow therein is uni-directional, where only motion coding provides motion vectors for frame coding. In this paper, we argue that, through information interactions, the synergy between motion coding and frame coding can be achieved. We effectively introduce bi-directional information interactions between motion coding and frame coding via our Motion Information Propagation. When generating the temporal contexts for frame coding, the high-dimension motion feature from the motion decoder serves as motion guidance to mitigate the alignment errors. Meanwhile, besides assisting frame coding at the current time step, the feature from context generation will be propagated as motion condition when coding the subsequent motion latent. Through the cycle of such interactions, feature propagation on motion coding is built, strengthening the capacity of exploiting long-range temporal correlation. In addition, we propose hybrid context generation to exploit the multi-scale context features and provide better motion condition. Experiments show that our method can achieve 12.9% bit rate saving over the previous SOTA neural video codec.
****
Accelerated Coordinate Encoding: Learning to Relocalize in Minutes Using RGB and Poses
Eric Brachmann, Tommaso Cavallari, Victor Adrian Prisacariu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5044-5053
Learning-based visual relocalizers exhibit leading pose accuracy, but require hours or days of training. Since training needs to happen on each new scene again, long training times make learning-based relocalization impractical for most applications, despite its promise of high accuracy. In this paper we show how such a system can actually achieve the same accuracy in less than 5 minutes. We start from the obvious: a relocalization network can be split in a scene-agnostic feature backbone, and a scene-specific prediction head. Less obvious: using an MLP prediction head allows us to optimize across thousands of view points simultaneously in each single training iteration. This leads to stable and extremely fast convergence. Furthermore, we substitute effective but slow end-to-end training using a robust pose solver with a curriculum over a reprojection loss. Our approach does not require privileged knowledge, such a depth maps or a 3D model, for speedy training. Overall, our approach is up to 300x faster in mapping than state-of-the-art scene coordinate regression, while keeping accuracy on par. Code is available: https://nianticlabs.github.io/ace
****
Switchable Representation Learning Framework With Self-Compatibility
Shengsen Wu, Yan Bai, Yihang Lou, Xiongkun Linghu, Jianzhong He, Ling-Yu Duan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15943-15953
Real-world visual search systems involve deployments on multiple platforms with different computing and storage resources. Deploying a unified model that suits the minimal-constrain platforms leads to limited accuracy. It is expected to deploy models with different capacities adapting to the resource constraints, which requires features extracted by these models to be aligned in the metric space. The method to achieve feature alignments is called "compatible learning". Existing research mainly focuses on the one-to-one compatible paradigm, which is limited in learning compatibility among multiple models. We propose a Switchable representation learning Framework with Self-Compatibility (SFSC). SFSC generates a series of compatible sub-models with different capacities through one training process. The optimization of sub-models faces gradients conflict, and we mitigate

this problem from the perspective of the magnitude and direction. We adjust the priorities of sub-models dynamically through uncertainty estimation to co-optimize sub-models properly. Besides, the gradients with conflicting directions are projected to avoid mutual interference. SFSC achieves state-of-the-art performance on the evaluated datasets.
****

Partial Network Cloning

Jingwen Ye, Songhua Liu, Xinchao Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 20137-20146

In this paper, we study a novel task that enables partial knowledge transfer from pre-trained models, which we term as Partial Network Cloning (PNC). Unlike prior methods that update all or at least part of the parameters in the target network throughout the knowledge transfer process, PNC conducts partial parametric "cloning" from a source network and then injects the cloned module to the target, without modifying its parameters. Thanks to the transferred module, the target network is expected to gain additional functionality, such as inference on new classes; whenever needed, the cloned module can be readily removed from the target, with its original parameters and competence kept intact. Specifically, we introduce an innovative learning scheme that allows us to identify simultaneously the component to be cloned from the source and the position to be inserted within the target network, so as to ensure the optimal performance. Experimental results on several datasets demonstrate that, our method yields a significant improvement of 5% in accuracy and 50% in locality when compared with parameter-tuning based methods.
****

MOTRv2: Bootstrapping End-to-End Multi-Object Tracking by Pretrained Object Detectors

Yuang Zhang, Tiancai Wang, Xiangyu Zhang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22056-22065

In this paper, we propose MOTRv2, a simple yet effective pipeline to bootstrap end-to-end multi-object tracking with a pretrained object detector. Existing end-to-end methods, e.g. MOTR and TrackFormer are inferior to their tracking-by-detection counterparts mainly due to their poor detection performance. We aim to improve MOTR by elegantly incorporating an extra object detector. We first adopt the anchor formulation of queries and then use an extra object detector to generate proposals as anchors, providing detection prior to MOTR. The simple modification greatly eases the conflict between joint learning detection and association tasks in MOTR. MOTRv2 keeps the end-to-end feature and scales well on large-scale benchmarks. MOTRv2 achieves the top performance (73.4% HOTA) among all existing methods on the DanceTrack dataset. Moreover, MOTRv2 reaches state-of-the-art performance on the BDD100K dataset. We hope this simple and effective pipeline can provide some new insights to the end-to-end MOT community. The code will be released in the near future.
****

Zero-Shot Dual-Lens Super-Resolution

Ruikang Xu, Mingde Yao, Zhiwei Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 9130-9139

The asymmetric dual-lens configuration is commonly available on mobile devices nowadays, which naturally stores a pair of wide-angle and telephoto images of the same scene to support realistic super-resolution (SR). Even on the same device, however, the degradation for modeling realistic SR is image-specific due to the unknown acquisition process (e.g., tiny camera motion). In this paper, we propose a zero-shot solution for dual-lens SR (ZeDuSR), where only the dual-lens pair at test time is used to learn an image-specific SR model. As such, ZeDuSR adapts itself to the current scene without using external training data, and thus gets rid of generalization difficulty. However, there are two major challenges to achieving this goal: 1) dual-lens alignment while keeping the realistic degradation, and 2) effective usage of highly limited training data. To overcome these two challenges, we propose a degradation-invariant alignment method and a degradation-aware training strategy to fully exploit the information within a single dua

l-lens pair. Extensive experiments validate the superiority of ZeDuSR over exist ing solutions on both synthesized and real-world dual-lens datasets.
****
Robust Dynamic Radiance Fields
Yu-Lun Liu, Chen Gao, Andréas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, Jia-Bin Huang; Proceedings of the IEEE/CVF Confe rence on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 13-23
Dynamic radiance field reconstruction methods aim to model the time-varying stru cture and appearance of a dynamic scene. Existing methods, however, assume that accurate camera poses can be reliably estimated by Structure from Motion (SfM) a lgorithms. These methods, thus, are unreliable as SfM algorithms often fail or p roduce erroneous poses on challenging videos with highly dynamic objects, poorly textured surfaces, and rotating camera motion. We address this issue by jointly estimating the static and dynamic radiance fields along with the camera paramet ers (poses and focal length). We demonstrate the robustness of our approach via extensive quantitative and qualitative experiments. Our results show favorable p erformance over the state-of-the-art dynamic view synthesis methods.
****
Improving Vision-and-Language Navigation by Generating Future-View Image Semanti cs
Jialu Li, Mohit Bansal; Proceedings of the IEEE/CVF Conference on Computer Visio n and Pattern Recognition (CVPR), 2023, pp. 10803-10812
Vision-and-Language Navigation (VLN) is the task that requires an agent to navig ate through the environment based on natural language instructions. At each step , the agent takes the next action by selecting from a set of navigable locations . In this paper, we aim to take one step further and explore whether the agent c an benefit from generating the potential future view during navigation. Intuitiv ely, humans will have an expectation of how the future environment will look lik e, based on the natural language instructions and surrounding views, which will aid correct navigation. Hence, to equip the agent with this ability to generate the semantics of future navigation views, we first propose three proxy tasks dur ing the agent's in-domain pre-training: Masked Panorama Modeling (MPM), Masked T rajectory Modeling (MTM), and Action Prediction with Image Generation (APIG). Th ese three objectives teach the model to predict missing views in a panorama (MPM), predict missing steps in the full trajectory (MTM), and generate the next vie w based on the full instruction and navigation history (APIG), respectively. We then fine-tune the agent on the VLN task with an auxiliary loss that minimizes t he difference between the view semantics generated by the agent and the ground t ruth view semantics of the next step. Empirically, our VLN-SIG achieves the new state-of-the-art on both the Room-to-Room dataset and the CVDN dataset. We furth er show that our agent learns to fill in missing patches in future views qualita tively, which brings more interpretability over agents' predicted actions. Lastl y, we demonstrate that learning to predict future view semantics also enables th e agent to have better performance on longer paths.
****
PLIKS: A Pseudo-Linear Inverse Kinematic Solver for 3D Human Body Estimation
Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, Bernhard Egger; Proceedings of the IEEE/CVF Conferen ce on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 574-584
We introduce PLIKS (Pseudo-Linear Inverse Kinematic Solver) for reconstruction o f a 3D mesh of the human body from a single 2D image. Current techniques directl y regress the shape, pose, and translation of a parametric model from an input i mage through a non-linear mapping with minimal flexibility to any external influ ences. We approach the task as a model-in-the-loop optimization problem. PLIKS i s built on a linearized formulation of the parametric SMPL model. Using PLIKS, w e can analytically reconstruct the human model via 2D pixel-aligned vertices. Th is enables us with the flexibility to use accurate camera calibration informatio n when available. PLIKS offers an easy way to introduce additional constraints s uch as shape and translation. We present quantitative evaluations which confirm that PLIKS achieves more accurate reconstruction with greater than 10% improveme

nt compared to other state-of-the-art methods with respect to the standard 3D human pose and shape benchmarks while also obtaining a reconstruction error improvement of 12.9 mm on the newer AGORA dataset.
****
Promoting Semantic Connectivity: Dual Nearest Neighbors Contrastive Learning for Unsupervised Domain Generalization
Yuchen Liu, Yaoming Wang, Yabo Chen, Wenrui Dai, Chenglin Li, Junni Zou, Hongkai Xiong; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3510-3519
Domain Generalization (DG) has achieved great success in generalizing knowledge from source domains to unseen target domains. However, current DG methods rely heavily on labeled source data, which are usually costly and unavailable. Since unlabeled data are far more accessible, we study a more practical unsupervised domain generalization (UDG) problem. Learning invariant visual representation from different views, i.e., contrastive learning, promises well semantic features for in-domain unsupervised learning. However, it fails in cross-domain scenarios. In this paper, we first delve into the failure of vanilla contrastive learning and point out that semantic connectivity is the key to UDG. Specifically, suppressing the intra-domain connectivity and encouraging the intra-class connectivity help to learn the domain-invariant semantic information. Then, we propose a novel unsupervised domain generalization approach, namely Dual Nearest Neighbors contrastive learning with strong Augmentation (DN^2A). Our DN^2A leverages strong augmentations to suppress the intra-domain connectivity and proposes a novel dual nearest neighbors search strategy to find trustworthy cross domain neighbors along with in-domain neighbors to encourage the intra-class connectivity. Experimental results demonstrate that our DN^2A outperforms the state-of-the-art by a large margin, e.g., 12.01% and 13.11% accuracy gain with only 1% labels for linear evaluation on PACS and DomainNet, respectively.
****
Interactive Segmentation of Radiance Fields
Rahul Goel, Dhawal Sirikonda, Saurabh Saini, P. J. Narayanan; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4201-4211
Radiance Fields (RF) are popular to represent casually-captured scenes for new view synthesis and several applications beyond it. Mixed reality on personal spaces needs understanding and manipulating scenes represented as RFs, with semantic segmentation of objects as an important step. Prior segmentation efforts show promise but don't scale to complex objects with diverse appearance. We present the ISRF method to interactively segment objects with fine structure and appearance. Nearest neighbor feature matching using distilled semantic features identifies high-confidence seed regions. Bilateral search in a joint spatio-semantic space grows the region to recover accurate segmentation. We show state-of-the-art results of segmenting objects from RFs and compositing them to another scene, changing appearance, etc., and an interactive segmentation tool that others can use.
****
gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction
Zerui Chen, Shizhe Chen, Cordelia Schmid, Ivan Laptev; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12890-12900
Signed distance functions (SDFs) is an attractive framework that has recently shown promising results for 3D shape reconstruction from images. SDFs seamlessly generalize to different shape resolutions and topologies but lack explicit modelling of the underlying 3D geometry. In this work, we exploit the hand structure and use it as guidance for SDF-based shape reconstruction. In particular, we address reconstruction of hands and manipulated objects from monocular RGB images. To this end, we estimate poses of hands and objects and use them to guide 3D reconstruction. More specifically, we predict kinematic chains of pose transformations and align SDFs with highly-articulated hand poses. We improve the visual features of 3D points with geometry alignment and further leverage temporal informat

ion to enhance the robustness to occlusion and motion blurs. We conduct extensive experiments on the challenging ObMan and DexYCB benchmarks and demonstrate significant improvements of the proposed method over the state of the art.
****
Principles of Forgetting in Domain-Incremental Semantic Segmentation in Adverse Weather Conditions
Tobias Kalb, Jürgen Beyerer; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19508-19518
Deep neural networks for scene perception in automated vehicles achieve excellent results for the domains they were trained on. However, in real-world conditions, the domain of operation and its underlying data distribution are subject to change. Adverse weather conditions, in particular, can significantly decrease model performance when such data are not available during training. Additionally, when a model is incrementally adapted to a new domain, it suffers from catastrophic forgetting, causing a significant drop in performance on previously observed domains. Despite recent progress in reducing catastrophic forgetting, its causes and effects remain obscure. Therefore, we study how the representations of semantic segmentation models are affected during domain-incremental learning in adverse weather conditions. Our experiments and representational analyses indicate that catastrophic forgetting is primarily caused by changes to low-level features in domain-incremental learning and that learning more general features on the source domain using pre-training and image augmentations leads to efficient feature reuse in subsequent tasks, which drastically reduces catastrophic forgetting. These findings highlight the importance of methods that facilitate generalized features for effective continual learning algorithms.
****
Neural Texture Synthesis With Guided Correspondence
Yang Zhou, Kaijian Chen, Rongjun Xiao, Hui Huang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 18095-18104
Markov random fields (MRFs) are the cornerstone of classical approaches to example-based texture synthesis. Yet, it is not fully valued in the deep learning era. This paper aims to re-promote the combination of MRFs and neural networks, i.e., the CNNMRF model, for texture synthesis, with two key observations made. We first propose to compute the Guided Correspondence Distance in the nearest neighbor search, based on which a Guided Correspondence loss is defined to measure the similarity of the output texture to the example. Experiments show that our approach surpasses existing neural approaches in uncontrolled and controlled texture synthesis. More importantly, the Guided Correspondence loss can function as a general textural loss in, e.g., training generative networks for real-time controlled synthesis and inversion-based single-image editing. In contrast, existing textural losses, such as the Sliced Wasserstein loss, cannot work on these challenging tasks.
****
Exploring and Utilizing Pattern Imbalance
Shibin Mei, Chenglong Zhao, Shengchao Yuan, Bingbing Ni; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7569-7578
In this paper, we identify pattern imbalance from several aspects, and further develop a new training scheme to avert pattern preference as well as spurious correlation. In contrast to prior methods which are mostly concerned with category or domain granularity, ignoring the potential finer structure that existed in datasets, we give a new definition of seed category as an appropriate optimization unit to distinguish different patterns in the same category or domain. Extensive experiments on domain generalization datasets of diverse scales demonstrate the effectiveness of the proposed method.
****
Are Data-Driven Explanations Robust Against Out-of-Distribution Data?
Tang Li, Fengchun Qiao, Mengmeng Ma, Xi Peng; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3821-3831

As black-box models increasingly power high-stakes applications, a variety of data-driven explanation methods have been introduced. Meanwhile, machine learning models are constantly challenged by distributional shifts. A question naturally arises: Are data-driven explanations robust against out-of-distribution data? Our empirical results show that even though predict correctly, the model might still yield unreliable explanations under distributional shifts. How to develop robust explanations against out-of-distribution data? To address this problem, we propose an end-to-end model-agnostic learning framework Distributionally Robust Explanations (DRE). The key idea is, inspired by self-supervised learning, to fully utilizes the inter-distribution information to provide supervisory signals for the learning of explanations without human annotation. Can robust explanations benefit the model's generalization capability? We conduct extensive experiments on a wide range of tasks and data types, including classification and regression on image and scientific tabular data. Our results demonstrate that the proposed method significantly improves the model's performance in terms of explanation and prediction robustness against distributional shifts.
****
Top-Down Visual Attention From Analysis by Synthesis
Baifeng Shi, Trevor Darrell, Xin Wang; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2102-2112
Current attention algorithms (e.g., self-attention) are stimulus-driven and highlight all the salient objects in an image. However, intelligent agents like humans often guide their attention based on the high-level task at hand, focusing only on task-related objects. This ability of task-guided top-down attention provides task-adaptive representation and helps the model generalize to various tasks. In this paper, we consider top-down attention from a classic Analysis-by-Synthesis (AbS) perspective of vision. Prior work indicates a functional equivalence between visual attention and sparse reconstruction; we show that an AbS visual system that optimizes a similar sparse reconstruction objective modulated by a goal-directed top-down signal naturally simulates top-down attention. We further propose Analysis-by-Synthesis Vision Transformer (AbSViT), which is a top-down modulated ViT model that variationally approximates AbS, and achieves controllable top-down attention. For real-world applications, AbSViT consistently improves over baselines on Vision-Language tasks such as VQA and zero-shot retrieval where language guides the top-down attention. AbSViT can also serve as a general backbone, improving performance on classification, semantic segmentation, and model robustness. Project page: https://sites.google.com/view/absvit.
****
Hierarchical Fine-Grained Image Forgery Detection and Localization
Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, Xiaoming Liu; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 3155-3165
Differences in forgery attributes of images generated in CNN-synthesized and image-editing domains are large, and such differences make a unified image forgery detection and localization (IFDL) challenging. To this end, we present a hierarchical fine-grained formulation for IFDL representation learning. Specifically, we first represent forgery attributes of a manipulated image with multiple labels at different levels. Then we perform fine-grained classification at these levels using the hierarchical dependency between them. As a result, the algorithm is encouraged to learn both comprehensive features and inherent hierarchical nature of different forgery attributes, thereby improving the IFDL representation. Our proposed IFDL framework contains three components: multi-branch feature extractor, localization and classification modules. Each branch of the feature extractor learns to classify forgery attributes at one level, while localization and classification modules segment the pixel-level forgery region and detect image-level forgery, respectively. Lastly, we construct a hierarchical fine-grained dataset to facilitate our study. We demonstrate the effectiveness of our method on 7 different benchmarks, for both tasks of IFDL and forgery attribute classification. Our source code and dataset can be found at https://github.com/CHELSEA234/HiFi_IFDL

****
CIMI4D: A Large Multimodal Climbing Motion Dataset Under Human-Scene Interactions
Ming Yan, Xin Wang, 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), 2023, pp. 12977-12988
Motion capture is a long-standing research problem. Although it has been studied for decades, the majority of research focus on ground-based movements such as walking, sitting, dancing, etc. Off-grounded actions such as climbing are largely overlooked. As an important type of action in sports and firefighting field, the climbing movements is challenging to capture because of its complex back poses, intricate human-scene interactions, and difficult global localization. The research community does not have an in-depth understanding of the climbing action due to the lack of specific datasets. To address this limitation, we collect CIMI4D, a large rock ClImbing MotIon on dataset from 12 persons climbing 13 different climbing walls. The dataset consists of around 180,000 frames of pose inertial measurements, LiDAR point clouds, RGB videos, high-precision static point cloud scenes, and reconstructed scene meshes. Moreover, we frame-wise annotate touch rock holds to facilitate a detailed exploration of human-scene interaction. The core of this dataset is a blending optimization process, which corrects for the pose as it drifts and is affected by the magnetic conditions. To evaluate the merit of CIMI4D, we perform four tasks which include human pose estimations (with/without scene constraints), pose prediction, and pose generation. The experimental results demonstrate that CIMI4D presents great challenges to existing methods and enables extensive research opportunities. We share the dataset with the research community in http://www.lidarhumanmotion.net/cimi4d/.
****
Fantastic Breaks: A Dataset of Paired 3D Scans of Real-World Broken Objects and Their Complete Counterparts
Nikolas Lamb, Cameron Palmer, Benjamin Molloy, Sean Banerjee, Natasha Kholgade Banerjee; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4681-4691
Automated shape repair approaches currently lack access to datasets that describe real-world damaged geometry. We present Fantastic Breaks (and Where to Find Them: https://terascale-all-sensing-research-studio.github.io/FantasticBreaks), a dataset containing scanned, waterproofed, and cleaned 3D meshes for 150 broken objects, paired and geometrically aligned with complete counterparts. Fantastic Breaks contains class and material labels, proxy repair parts that join to broken meshes to generate complete meshes, and manually annotated fracture boundaries. Through a detailed analysis of fracture geometry, we reveal differences between Fantastic Breaks and synthetic fracture datasets generated using geometric and physics-based methods. We show experimental shape repair evaluation with Fantastic Breaks using multiple learning-based approaches pre-trained with synthetic datasets and re-trained with subset of Fantastic Breaks.
****
Modernizing Old Photos Using Multiple References via Photorealistic Style Transfer
Agus Gunawan, Soo Ye Kim, Hyeonjun Sim, Jae-Ho Lee, Munchurl Kim; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 12460-12469
This paper firstly presents old photo modernization using multiple references by performing stylization and enhancement in a unified manner. In order to modernize old photos, we propose a novel multi-reference-based old photo modernization (MROPM) framework consisting of a network MROPM-Net and a novel synthetic data generation scheme. MROPM-Net stylizes old photos using multiple references via photorealistic style transfer (PST) and further enhances the results to produce modern-looking images. Meanwhile, the synthetic data generation scheme trains the network to effectively utilize multiple references to perform modernization. To evaluate the performance, we propose a new old photos benchmark dataset (CHD) consisting of diverse natural indoor and outdoor scenes. Extensive experiments sho

w that the proposed method outperforms other baselines in performing modernizati on on real old photos, even though no old photos were used during training. More over, our method can appropriately select styles from multiple references for ea ch semantic region in the old photo to further improve the modernization perform ance.
****

Interactive Cartoonization With Controllable Perceptual Factors
Namhyuk Ahn, Patrick Kwon, Jihye Back, Kibeom Hong, Seungkwon Kim; Proceedings o f the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 202 3, pp. 16827-16835
Cartoonization is a task that renders natural photos into cartoon styles. Previo us deep methods only have focused on end-to-end translation, disabling artists f rom manipulating results. To tackle this, in this work, we propose a novel solut ion with editing features of texture and color based on the cartoon creation pro cess. To do that, we design a model architecture to have separate decoders, text ure and color, to decouple these attributes. In the texture decoder, we propose a texture controller, which enables a user to control stroke style and abstracti on to generate diverse cartoon textures. We also introduce an HSV color augmenta tion to induce the networks to generate consistent color translation. To the bes t of our knowledge, our work is the first method to control the cartoonization d uring the inferences step, generating high-quality results compared to baselines .
****

Curvature-Balanced Feature Manifold Learning for Long-Tailed Classification
Yanbiao Ma, Licheng Jiao, Fang Liu, Shuyuan Yang, Xu Liu, Lingling Li; Proceedin gs of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 15824-15835
To address the challenges of long-tailed classification, researchers have propos ed several approaches to reduce model bias, most of which assume that classes wi th few samples are weak classes. However, recent studies have shown that tail cl asses are not always hard to learn, and model bias has been observed on sample-b alanced datasets, suggesting the existence of other factors that affect model bi as. In this work, we systematically propose a series of geometric measures for p erceptual manifolds in deep neural networks, and then explore the effect of the geometric characteristics of perceptual manifolds on classification difficulty a nd how learning shapes the geometric characteristics of perceptual manifolds. An unanticipated finding is that the correlation between the class accuracy and th e separation degree of perceptual manifolds gradually decreases during training, while the negative correlation with the curvature gradually increases, implying that curvature imbalance leads to model bias. Therefore, we propose curvature r egularization to facilitate the model to learn curvature-balanced and flatter pe rceptual manifolds. Evaluations on multiple long-tailed and non-long-tailed data sets show the excellent performance and exciting generality of our approach, esp ecially in achieving significant performance improvements based on current state -of-the-art techniques. Our work reminds researchers to pay attention to model b ias not only on long-tailed datasets but also on non-long-tailed and even data-b alanced datasets, which can improve model performance from another perspective.
****