Towards Attack-tolerant Federated Learning via Critical Parameter Analysis
Sungwon Han, Sungwon Park, Fangzhao Wu, Sundong Kim, Bin Zhu, Xing Xie, Meeyoung Cha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4999-5008
Federated learning is used to train a shared model in a decentralized way without clients sharing private data with each other. Federated learning systems are susceptible to poisoning attacks when malicious clients send false updates to the central server. Existing defense strategies are ineffective under non-IID data settings. This paper proposes a new defense strategy, FedCPA (Federated learning with Critical Parameter Analysis). Our attack-tolerant aggregation method is based on the observation that benign local models have similar sets of top-k and bottom-k critical parameters, whereas poisoned local models do not. Experiments with different attack scenarios on multiple datasets demonstrate that our model outperforms existing defense strategies in defending against poisoning attacks.
****
Stochastic Segmentation with Conditional Categorical Diffusion Models
Lukas Zbinden, Lars Doorenbos, Theodoros Pissas, Adrian Thomas Huber, Raphael Sznitman, Pablo Márquez-Neila; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1119-1129
Semantic segmentation has made significant progress in recent years thanks to deep neural networks, but the common objective of generating a single segmentation output that accurately matches the image's content may not be suitable for safety-critical domains such as medical diagnostics and autonomous driving. Instead, multiple possible correct segmentation maps may be required to reflect the true distribution of annotation maps. In this context, stochastic semantic segmentation methods must learn to predict conditional distributions of labels given the image, but this is challenging due to the typically multimodal distributions, high-dimensional output spaces, and limited annotation data. To address these challenges, we propose a conditional categorical diffusion model (CCDM) for semantic segmentation based on Denoising Diffusion Probabilistic Models. Our model is conditioned to the input image, enabling it to generate multiple segmentation label maps that account for the aleatoric uncertainty arising from divergent ground truth annotations. Our experimental results show that CCDM achieves state-of-the-art performance on LIDC, a stochastic semantic segmentation dataset, and outperforms established baselines on the classical segmentation dataset Cityscapes.
****
Diff-Retinex: Rethinking Low-light Image Enhancement with A Generative Diffusion Model
Xunpeng Yi, Han Xu, Hao Zhang, Linfeng Tang, Jiayi Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12302-12311
In this paper, we rethink the low-light image enhancement task and propose a physically explainable and generative diffusion model for low-light image enhancement, termed as Diff-Retinex. We aim to integrate the advantages of the physical model and the generative network. Furthermore, we hope to supplement and even deduce the information missing in the low-light image through the generative network. Therefore, Diff-Retinex formulates the low-light image enhancement problem into Retinex decomposition and conditional image generation. In the Retinex decomposition, we integrate the superiority of attention in Transformer and meticulously design a Retinex Transformer decomposition network (TDN) to decompose the image into illumination and reflectance maps. Then, we design multi-path generative diffusion networks to reconstruct the normal-light Retinex probability distribution and solve the various degradations in these components respectively, including dark illumination, noise, color deviation, loss of scene contents, etc. Owing to generative diffusion model, Diff-Retinex puts the restoration of low-light subtle detail into practice. Extensive experiments conducted on real-world low-light datasets qualitatively and quantitatively demonstrate the effectiveness, superiority, and generalization of the proposed method.
****
Bird's-Eye-View Scene Graph for Vision-Language Navigation
Rui Liu, Xiaohan Wang, Wenguan Wang, Yi Yang; Proceedings of the IEEE/CVF Intern

ational Conference on Computer Vision (ICCV), 2023, pp. 10968-10980
Vision-language navigation (VLN), which entails an agent to navigate 3D environm
ents following human instructions, has shown great advances. However, current ag
ents are built upon panoramic observations, which hinders their ability to perce
ive 3D scene geometry and easily leads to ambiguous selection of panoramic view.
 To address these limitations, we present a BEV Scene Graph (BSG), which leverag
es multi-step BEV representations to encode scene layouts and geometric cues of
indoor environment under the supervision of 3D detection. During navigation, BSG
 builds a local BEV representation at each step and maintains a BEV-based global
 scene map, which stores and organizes all the online collected local BEV repres
entations according to their topological relations. Based on BSG, the agent pred
icts a local BEV grid-level decision score and a global graph-level decision sco
re, combined with a subview selection score on panoramic views, for more accurat
e action prediction. Our approach significantly outperforms state-of-the-art met
hods on REVERIE, R2R, and R4R, showing the potential of BEV perception in VLN.
****
PVT++: A Simple End-to-End Latency-Aware Visual Tracking Framework
Bowen Li, Ziyuan Huang, Junjie Ye, Yiming Li, Sebastian Scherer, Hang Zhao, Chan
ghong Fu; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2023, pp. 10006-10016
Visual object tracking is essential to intelligent robots. Most existing approac
hes have ignored the online latency that can cause severe performance degradatio
n during real-world processing. Especially for unmanned aerial vehicles (UAVs),
where robust tracking is more challenging and onboard computation is limited, th
e latency issue can be fatal. In this work, we present a simple framework for en
d-to-end latency-aware tracking, i.e., end-to-end predictive visual tracking (PV
T++). Unlike existing solutions that naively append Kalman Filters after tracker
s, PVT++ can be jointly optimized, so that it takes not only motion information
but can also leverage the rich visual knowledge in most pre-trained tracker mode
ls for robust prediction. Besides, to bridge the training-evaluation domain gap,
 we propose a relative motion factor, empowering PVT++ to generalize to the chal
lenging and complex UAV tracking scenes. These careful designs have made the sma
ll-capacity lightweight PVT++ a widely effective solution. Additionally, this wo
rk presents an extended latency-aware evaluation benchmark for assessing an any-
speed tracker in the online setting. Empirical results on a robotic platform fro
m the aerial perspective show that PVT++ can achieve significant performance gai
n on various trackers and exhibit higher accuracy than prior solutions, largely
mitigating the degradation brought by latency. Our code will be made public.
****
A Dynamic Dual-Processing Object Detection Framework Inspired by the Brain's Rec
ognition Mechanism
Minying Zhang, Tianpeng Bu, Lulu Hu; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 6264-6274
There are two main approaches to object detection: CNN-based and Transformer-bas
ed. The former views object detection as a dense local matching problem, while t
he latter sees it as a sparse global retrieval problem. Research in neuroscience
 has shown that the recognition decision in the brain is based on two processes,
 namely familiarity and recollection. Based on this biological support, we propo
se an efficient and effective dual-processing object detection framework. It int
egrates CNN- and Transformer-based detectors into a comprehensive object detecti
on system consisting of a shared backbone, an efficient dual-stream encoder, and
 a dynamic dual-decoder. To better integrate local and global features, we desig
n a search space for the CNN-Transformer dual-stream encoder to find the optimal
 fusion solution. To enable better coordination between the CNN- and Transformer
-based decoders, we provide the dual-decoder with a selective mask. This mask dy
namically chooses the more advantageous decoder for each position in the image b
ased on high-level representation. As demonstrated by extensive experiments, our
 approach shows flexibility and effectiveness in prompting the mAP of the variou
s source detectors by 3.0 3.7 without increasing FLOPs.
****

Hard No-Box Adversarial Attack on Skeleton-Based Human Action Recognition with Skeleton-Motion-Informed Gradient

Zhengzhi Lu, He Wang, Ziyi Chang, Guoan Yang, Hubert P. H. Shum; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4597-4606

Recently, methods for skeleton-based human activity recognition have been shown to be vulnerable to adversarial attacks. However, these attack methods require either the full knowledge of the victim (i.e. white-box attacks), access to training data (i.e. transfer-based attacks) or frequent model queries (i.e. black-box attacks). All their requirements are highly restrictive, raising the question of how detrimental the vulnerability is. In this paper, we show that the vulnerability indeed exists. To this end, we consider a new attack task: the attacker has no access to the victim model or the training data or labels, where we coin the term hard no-box attack. Specifically, we first learn a motion manifold where we define an adversarial loss to compute a new gradient for the attack, named skeleton-motion-informed (SMI) gradient. Our gradient contains information of the motion dynamics, which is different from existing gradient-based attack methods that compute the loss gradient assuming each dimension in the data is independent. The SMI gradient can augment many gradient-based attack methods, leading to a new family of no-box attack methods. Extensive evaluation and comparison show that our method imposes a real threat to existing classifiers. They also show that the SMI gradient improves the transferability and imperceptibility of adversarial samples in both no-box and transfer-based black-box settings.

****

GameFormer: Game-theoretic Modeling and Learning of Transformer-based Interactive Prediction and Planning for Autonomous Driving

Zhiyu Huang, Haochen Liu, Chen Lv; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3903-3913

Autonomous vehicles operating in complex real-world environments require accurate predictions of interactive behaviors between traffic participants. This paper tackles the interaction prediction problem by formulating it with hierarchical game theory and proposing the GameFormer model for its implementation. The model incorporates a Transformer encoder, which effectively models the relationships between scene elements, alongside a novel hierarchical Transformer decoder structure. At each decoding level, the decoder utilizes the prediction outcomes from the previous level, in addition to the shared environmental context, to iteratively refine the interaction process. Moreover, we propose a learning process that regulates an agent's behavior at the current level to respond to other agents' behaviors from the preceding level. Through comprehensive experiments on large-scale real-world driving datasets, we demonstrate the state-of-the-art accuracy of our model on the Waymo interaction prediction task. Additionally, we validate the model's capacity to jointly reason about the motion plan of the ego agent and the behaviors of multiple agents in both open-loop and closed-loop planning tests, outperforming various baseline methods. Furthermore, we evaluate the efficacy of our model on the nuPlan planning benchmark, where it achieves leading performance.

****

Towards Better Robustness against Common Corruptions for Unsupervised Domain Adaptation

Zhiqiang Gao, Kaizhu Huang, Rui Zhang, Dawei Liu, Jieming Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18882-18893

Recent studies have investigated how to achieve robustness for unsupervised domain adaptation (UDA). While most efforts focus on adversarial robustness, i.e. how the model performs against unseen malicious adversarial perturbations, robustness against benign common corruption (RaCC) surprisingly remains under-explored for UDA. Towards improving RaCC for UDA methods in an unsupervised manner, we propose a novel Distributionally and Discretely Adversarial Regularization (DDAR) framework in this paper. Formulated as a min-max optimization with a distribution distance, DDAR is theoretically well-founded to ensure generalization over unk

nown common corruptions. Meanwhile, we show that our regularization scheme effec tively reduces a surrogate of RaCC, i.e., the perceptual distance between natura l data and common corruption. To enable a better adversarial regularization, the design of the optimization pipeline relies on an image discretization scheme th at can transform "out-of-distribution" adversarial data into "in-distribution" d ata augmentation. Through extensive experiments, in terms of RaCC, our method is superior to conventional unsupervised regularization mechanisms, widely improve s the robustness of existing UDA methods, and achieves state-of-the-art performa nce.
****
Learning in Imperfect Environment: Multi-Label Classification with Long-Tailed D istribution and Partial Labels
Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, Yueting Z huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1423-1432
Conventional multi-label classification (MLC) methods assume that all samples ar e fully labeled and identically distributed. Unfortunately, this assumption is u nrealistic in large-scale MLC data
that has long-tailed (LT) distribution and partial labels (PL).
To address the problem, we introduce a novel task, Partial labeling and Long-Ta iled Multi-Label Classification (PLT-MLC), to jointly consider the above two imp erfect learning environments. Not surprisingly, we find that most LT-MLC and PL- MLC approaches fail to solve the PLT-MLC, resulting in significant performance d egradation on the two proposed PLT-MLC benchmarks. Therefore, we propose an end- to-end learning framework: COrrection -> ModificatIon -> balanCe, abbreviated as COMC.
Our bootstrapping philosophy is to simultaneously correct the missing labels (C orrection) with convinced prediction confidence over a class-aware threshold and to learn from these recall labels during training.
We next propose a novel multi-focal modifier loss that simultaneously addresses head-tail imbalance and positive-negative imbalance to adaptively modify the at tention to different samples (Modification) under the LT class distribution.
We also develop a balanced training strategy by distilling the model's learning effect from head and tail samples, and thus design the balanced classifier (Bal ance) conditioned on the head and tail learning effect to maintain a stable perf ormance.
Our experimental study shows that the proposed method significantly outperforms the general MLC, LT-MLC and ML-MLC methods in terms of effectiveness and robust ness on our newly created PLT-MLC datasets.
****
Flexible Visual Recognition by Evidential Modeling of Confusion and Ignorance
Lei Fan, Bo Liu, Haoxiang Li, Ying Wu, Gang Hua; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2023, pp. 1338-1347
In real-world scenarios, typical visual recognition systems could fail under two major causes, i.e., the misclassification between known classes and the excusab le misbehavior on unknown-class images. To tackle these deficiencies, flexible v isual recognition should dynamically predict multiple classes when they are unco nfident between choices and reject making predictions when the input is entirely out of the training distribution. Two challenges emerge along with this novel t ask. First, prediction uncertainty should be separately quantified as confusion depicting inter-class uncertainties and ignorance identifying out-of-distributio n samples. Second, both confusion and ignorance should be comparable between sam ples to enable effective decision-making. In this paper, we propose to model the se two sources of uncertainty explicitly with the theory of Subjective Logic. Re garding recognition as an evidence-collecting process, confusion is then defined as conflicting evidence, while ignorance is the absence of evidence. By predict ing Dirichlet concentration parameters for singletons, comprehensive subjective opinions, including confusion and ignorance, could be achieved via further evide nce combinations. Through a series of experiments on synthetic data analysis, vi sual recognition, and open-set detection, we demonstrate the effectiveness of ou

r methods in quantifying two sources of uncertainties and dealing with flexible recognition.
****
Texture Generation on 3D Meshes with Point-UV Diffusion
Xin Yu, Peng Dai, Wenbo Li, Lan Ma, Zhengzhe Liu, Xiaojuan Qi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4206-4216
In this work, we focus on synthesizing high-quality textures on 3D meshes. We present Point-UV diffusion, a coarse-to-fine pipeline that marries the denoising diffusion model with UV mapping to generate 3D consistent and high-quality texture images in UV space. We start with introducing a point diffusion model to synthesize low-frequency texture components with our tailored style guidance to tackle the biased color distribution. The derived coarse texture offers global consistency and serves as a condition for the subsequent UV diffusion stage, aiding in regularizing the model to generate a 3D consistent UV texture image. Then, a UV diffusion model with hybrid conditions is developed to enhance the texture fidelity in the 2D UV space. Our method can process meshes of any genus, generating diversified, geometry-compatible, and high-fidelity textures.
****
Supervised Homography Learning with Realistic Dataset Generation
Hai Jiang, Haipeng Li, Songchen Han, Haoqiang Fan, Bing Zeng, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9806-9815
In this paper, we propose an iterative framework, which consists of two phases: a generation phase and a training phase, to generate realistic training data and yield a supervised homography network. In the generation phase, given an unlabeled image pair, we utilize the pre-estimated dominant plane masks and homography of the pair, along with another sampled homography that serves as ground truth to generate a new labeled training pair with realistic motion. In the training phase, the generated data is used to train the supervised homography network, in which the training data is refined via a content consistency module and a quality assessment module. Once an iteration is finished, the trained network is used in the next data generation phase to update the pre-estimated homography. Through such an iterative strategy, the quality of the dataset and the performance of the network can be gradually and simultaneously improved. Experimental results show that our method achieves state-of-the-art performance and existing supervised methods can be also improved based on the generated dataset. Code and dataset are available at https://github.com/JianghaiSCU/RealSH.
****
E2E-LOAD: End-to-End Long-form Online Action Detection
Shuqiang Cao, Weixin Luo, Bairui Wang, Wei Zhang, Lin Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10422-10432
Recently, feature-based methods for Online Action Detection (OAD) have been gaining traction. However, these methods are constrained by their fixed backbone design, which fails to leverage the potential benefits of a trainable backbone. This paper introduces an end-to-end learning network that revises these approaches, incorporating a backbone network design that improves effectiveness and efficiency. Our proposed model utilizes a shared initial spatial model for all frames and maintains an extended sequence cache, which enables low-cost inference. We promote an asymmetric spatiotemporal model that caters to long-form and short-form modeling. Additionally, we propose an innovative and efficient inference mechanism that accelerates extensive spatiotemporal exploration. Through comprehensive ablation studies and experiments, we validate the performance and efficiency of our proposed method. Remarkably, we achieve an end-to-end learning OAD of 17.3 (+12.6) FPS with 72.4% (+1.2%), 90.3% (+0.7%), and 48.1% (+26.0%) mAP on THMOUS'14, TVSeries, and HDD, respectively.
****
TALL: Thumbnail Layout for Deepfake Video Detection
Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, Ran He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.

 22658-22668
The growing threats of deepfakes to society and cybersecurity have raised enormo
us public concerns, and increasing efforts have been devoted to this critical to
pic of deepfake video detection. Existing video methods achieve good performance
 but are computationally intensive. This paper introduces a simple yet effective
 strategy named Thumbnail Layout (TALL), which transforms a video clip into a pr
e-defined layout to realize the preservation of spatial and temporal dependencie
s. Specifically, consecutive frames are masked in a fixed position in each frame
 to improve generalization, then resized to sub-images and rearranged into a pre
-defined layout as the thumbnail. TALL is model-agnostic and extremely simple by
 only modifying a few lines of code. Inspired by the success of vision transform
ers, we incorporate TALL into Swin Transformer, forming an efficient and effecti
ve method TALL-Swin. Extensive experiments on intra-dataset and cross-dataset va
lidate the validity and superiority of TALL and SOTA TALL-Swin. TALL-Swin achiev
es 90.79% AUC on the challenging cross-dataset task, FaceForensics++ - Celeb-DF.
 The code is available at https://github.com/rainy-xu/TALL4Deepfake.
****
Enhanced Soft Label for Semi-Supervised Semantic Segmentation
Jie Ma, Chuan Wang, Yang Liu, Liang Lin, Guanbin Li; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2023, pp. 1185-1195
As a mainstream framework in the field of semi-supervised learning (SSL), self-t
raining via pseudo labeling and its variants have witnessed impressive progress
in semi-supervised semantic segmentation with the recent advance of deep neural
networks. However, modern self-training based SSL algorithms use a pre-defined c
onstant threshold to select unlabeled pixel samples that contribute to the train
ing, thus failing to be compatible with different learning difficulties of varia
nt categories and different learning status of the model. To address these issue
s, we propose Enhanced Soft Label (ESL), a curriculum learning approach to fully
 leverage the high-value supervisory signals implicit in the untrustworthy pseud
o label. ESL believes that pixels with unconfident predictions can be pretty sur
e about their belonging to a subset of dominant classes though being arduous to
determine the exact one. It thus contains a Dynamic Soft Label (DSL) module to d
ynamically maintain the high probability classes, keeping the label "soft" so as
 to make full use of the high entropy prediction. However, the DSL itself will i
nevitably introduce ambiguity between dominant classes, thus blurring the classi
fication boundary. Therefore, we further propose a pixel-to-part contrastive lea
rning method cooperated with an unsupervised object part grouping mechanism to i
mprove its ability to distinguish between different classes. Extensive experimen
tal results on Pascal VOC 2012 and Cityscapes show that our approach achieves re
markable improvements over existing state-of-the-art approaches.
****
Self-supervised Monocular Depth Estimation: Let's Talk About The Weather
Kieran Saunders, George Vogiatzis, Luis J. Manso; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2023, pp. 8907-8917
Current, self-supervised depth estimation architectures rely on clear and sunny
weather scenes to train deep neural networks. However, in many locations, this a
ssumption is too strong. For example in the UK (2021), 149 days consisted of rai
n. For these architectures to be effective in real-world applications, we must c
reate models that can generalise to all weather conditions, times of the day and
 image qualities. Using a combination of computer graphics and generative models
, one can augment existing sunny-weather data in a variety of ways that simulate
 adverse weather effects. While it is tempting to use such data augmentations fo
r self-supervised depth, in the past this was shown to degrade performance inste
ad of improving it. In this paper, we put forward a method that uses augmentatio
ns to remedy this problem. By exploiting the correspondence between unaugmented
and augmented data we introduce a pseudo-supervised loss for both depth and pose
 estimation. This brings back some of the benefits of supervised learning while
still not requiring any labels. We also make a series of practical recommendatio
ns which collectively offer a reliable, efficient framework for weather-related
augmentation of self-supervised depth from monocular video. We present extensive

testing to show that our method, Robust-Depth, achieves SotA performance on the KITTI dataset while significantly surpassing SotA on challenging, adverse condition data such as DrivingStereo, Foggy CityScape and NuScenes-Night. The project website can be found at https://kieran514.github.io/Robust-Depth-Project/.
****
Bidirectional Alignment for Domain Adaptive Detection with Transformers
Liqiang He, Wei Wang, Albert Chen, Min Sun, Cheng-Hao Kuo, Sinisa Todorovic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18775-18785
We propose a Bidirectional Alignment for domain adaptive Detection with Transformers (BiADT) to improve cross domain object detection performance. Existing adversarial learning based methods use gradient reverse layer (GRL) to reduce the domain gap between the source and target domains in feature representations. Since different image parts and objects may exhibit various degrees of domain-specific characteristics, directly applying GRL on a global image or object representation may not be suitable. Our proposed BiADT explicitly estimates token-wise domain-invariant and domain-specific features in the image and object token sequences. BiADT has a novel deformable attention and self-attention, aimed at bi-directional domain alignment and mutual information minimization. These two objectives reduce the domain gap in domain-invariant representations, and simultaneously increase the distinctiveness of domain-specific features. Our experiments show that BiADT achieves very competitive performance to SOTA consistently on Cityscapes-to-FoggyCityscapes, Sim10K-to-Citiscapes and Cityscapes-to-BDD100K, outperforming the strong baseline, AQT, by 2.0, 2.1, and 2.4 in mAP50, respectively.
****
Fast Neural Scene Flow
Xueqian Li, Jianqiao Zheng, Francesco Ferroni, Jhony Kaesemodel Pontes, Simon Lucey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9878-9890
Neural Scene Flow Prior (NSFP) is of significant interest to the vision community due to its inherent robustness to out-of-distribution (OOD) effects and its ability to deal with dense lidar points. The approach utilizes a coordinate neural network to estimate scene flow at runtime, without any training. However, it is up to 100 times slower than current state-of-the-art learning methods. In other applications such as image, video, and radiance function reconstruction innovations in speeding up the runtime performance of coordinate networks have centered upon architectural changes. In this paper, we demonstrate that scene flow is different---with the dominant computational bottleneck stemming from the loss function itself (i.e., Chamfer distance). Further, we rediscover the distance transform (DT) as an efficient, correspondence-free loss function that dramatically speeds up the runtime optimization. Our fast neural scene flow (FNSF) approach reports for the first time real-time performance comparable to learning methods, without any training or OOD bias on two of the largest open autonomous driving (AV) lidar datasets Waymo Open [62] and Argoverse [8].
****
CAME: Contrastive Automated Model Evaluation
Ru Peng, Qiuyang Duan, Haobo Wang, Jiachen Ma, Yanbo Jiang, Yongjun Tu, Xiu Jiang, Junbo Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20121-20132
The Automated Model Evaluation (AutoEval) framework entertains the possibility of evaluating a trained machine learning model without resorting to a labeled testing set.
Despite the promise and some decent results, the existing AutoEval methods heavily rely on computing distribution shifts between the unlabelled testing set and the training set.
We believe this reliance on the training set becomes another obstacle in shipping this technology to real-world ML development.
In this work, we propose Contrastive Automatic Model Evaluation (CAME), a novel AutoEval framework that is rid of involving training set in the loop.
The core idea of CAME bases on a theoretical analysis which bonds the model per

formance with a contrastive loss.
Further, with extensive empirical validation, we manage to set up a predictable relationship between the two, simply by deducing on the unlabeled/unseen testing set.
The resulting framework CAME establishes a new SOTA results for AutoEval by surpassing prior work significantly.
****
ExposureDiffusion: Learning to Expose for Low-light Image Enhancement
Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C. Kot, Bihan Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12438-12448
Previous raw image-based low-light image enhancement methods predominantly relied on feed-forward neural networks to learn deterministic mappings from low-light to normally-exposed images. However, they failed to capture critical distribution information, leading to visually undesirable results. This work addresses the issue by seamlessly integrating a diffusion model with a physics-based exposure model. Different from a vanilla diffusion model that has to perform Gaussian denoising, with the injected physics-based exposure model, our restoration process can directly start from a noisy image instead of pure noise. As such, our method obtains significantly improved performance and reduced inference time compared with vanilla diffusion models. To make full use of the advantages of different intermediate steps, we further propose an adaptive residual layer that effectively screens out the side-effect in the iterative refinement when the intermediate results have been already well-exposed. The proposed framework can work with both real-paired datasets, SOTA noise models, and different backbone networks. We evaluate the proposed method on various public benchmarks, achieving promising results with consistent improvements using different exposure models and backbones. Besides, the proposed method achieves better generalization capacity for unseen amplifying ratios and better performance than a larger feedforward neural model when few parameters are adopted. The code is released at https://github.com/wyf0912/ExposureDiffusion.
****
HM-ViT: Hetero-Modal Vehicle-to-Vehicle Cooperative Perception with Vision Transformer
Hao Xiang, Runsheng Xu, Jiaqi Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 284-295
Vehicle-to-Vehicle technologies have enabled autonomous vehicles to share information to see through occlusions, greatly enhancing perception performance. Nevertheless, existing works all focused on homogeneous traffic where vehicles are equipped with the same type of sensors, which significantly hampers the scale of collaboration and benefit of cross-modality interactions. In this paper, we investigate the multi-agent hetero-modal cooperative perception problem where agents may have distinct sensor modalities. We present HM-ViT, the first unified multi-agent hetero-modal cooperative perception framework that can collaboratively predict 3D objects for highly dynamic Vehicle-to-Vehicle (V2V) collaborations with varying numbers and types of agents. To effectively fuse features from multi-view images and LiDAR point clouds, we design a novel heterogeneous 3D graph transformer to jointly reason inter-agent and intra-agent interactions. The extensive experiments on the V2V perception dataset OPV2V demonstrate that the HM-ViT outperforms SOTA cooperative perception methods for V2V hetero-modal cooperative perception. Our code will be released at https://github.com/XHwind/HM-ViT.
****
HyperReenact: One-Shot Reenactment via Jointly Learning to Refine and Retarget Faces
Stella Bounareli, Christos Tzelepis, Vasileios Argyriou, Ioannis Patras, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7149-7159
In this paper, we present our method for neural face reenactment, called HyperReenact, that aims to generate realistic talking head images of a source identity, driven by a target facial pose. Existing state-of-the-art face reenactment meth

ods train controllable generative models that learn to synthesize realistic facial images, yet producing reenacted faces that are prone to significant visual artifacts, especially under the challenging condition of extreme head pose changes, or requiring expensive few-shot fine-tuning to better preserve the source identity characteristics. We propose to address these limitations by leveraging the photorealistic generation ability and the disentangled properties of a pretrained StyleGAN2 generator, by first inverting the real images into its latent space and then using a hypernetwork to perform: (i) refinement of the source identity characteristics and (ii) facial pose re-targeting, eliminating this way the dependence on external editing methods that typically produce artifacts. Our method operates under the one-shot setting (i.e., using a single source frame) and allows for cross-subject reenactment, without requiring any subject-specific fine-tuning. We compare our method both quantitatively and qualitatively against several state-of-the-art techniques on the standard benchmarks of VoxCeleb1 and VoxCeleb2, demonstrating the superiority of our approach in producing artifact-free images, exhibiting remarkable robustness even under extreme head pose changes. We make the code and the pretrained models publicly available at: https://github.com/StelaBou/HyperReenact
****
Order-preserving Consistency Regularization for Domain Adaptation and Generalization
Mengmeng Jing, Xiantong Zhen, Jingjing Li, Cees G. M. Snoek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18916-18927
Deep learning models fail on cross-domain challenges if the model is oversensitive to domain-specific attributes, e.g., lightning, background, camera angle, etc. To alleviate this problem, data augmentation coupled with consistency regularization are commonly adopted to make the model less sensitive to domain-specific attributes. Consistency regularization enforces the model to output the same representation or prediction for two views of one image. These constraints, however, are either too strict or not order-preserving for the classification probabilities. In this work, we propose the Order-preserving Consistency Regularization (OCR) for cross-domain tasks. The order-preserving property for the prediction makes the model robust to task-irrelevant transformations. As a result, the model becomes less sensitive to the domain-specific attributes. The comprehensive experiments show that our method achieves clear advantages on five different cross-domain tasks.
****
RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D
Shuhei Kurita, Naoki Katsura, Eri Onami; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15214-15224
Grounding textual expressions on scene objects from first-person views is a truly demanding capability in developing agents that are aware of their surroundings and behave following intuitive text instructions. Such capability is of necessity for glass-devices or autonomous robots to localize referred objects in the real-world. In the conventional referring expression comprehension tasks of images, however, datasets are mostly constructed based on the web-crawled data and don't reflect diverse real-world structures on the task of grounding textual expressions in diverse objects in the real world. Recently, a massive-scale egocentric video dataset of Ego4D was proposed. Ego4D covers around the world diverse real-world scenes including numerous indoor and outdoor situations such as shopping, cooking, walking, talking, manufacturing, etc. Based on egocentric videos of Ego4D, we constructed a broad coverage of the video-based referring expression comprehension dataset: RefEgo. Our dataset includes more than 12k video clips and 41 hours for video-based referring expression comprehension annotation.
In experiments, we combine the state-of-the-art 2D referring expression comprehension models with the object tracking algorithm, achieving the video-wise referred object tracking even in difficult conditions: the referred object becomes out-of-frame in the middle of the video or multiple similar objects are presented

in the video.
****

Exploring Temporal Frequency Spectrum in Deep Video Deblurring
Qi Zhu, Man Zhou, Naishan Zheng, Chongyi Li, Jie Huang, Feng Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12428-12437
Video deblurring aims to restore the latent video frames from their blurred counterparts. Despite the remarkable progress, most promising video deblurring methods only investigate the temporal priors in the spatial domain and rarely explore their its potential in the frequency domain. In this paper, we revisit the blurred sequence in the Fourier space and figure out some intrinsic frequency-temporal priors that imply the temporal blur degradation can be accessibly decoupled in the potential frequency domain. Based on these priors, we propose a novel Fourier-based frequency-temporal video deblurring solution, where the core design accommodates the temporal spectrum to a popular video deblurring pipeline of feature extraction, alignment, aggregation, and optimization.
Specifically, we design a Spectrum Prior-guided Alignment module by leveraging enlarged blur information in the potential spectrum to mitigate the blur effects on the alignment. Then, Temporal Energy prior-driven Aggregation is implemented to replenish the original local features by estimating the temporal spectrum energy as the global sharpness guidance. In addition, the customized frequency loss is devised to optimize the proposed method for decent spectral distribution.
Extensive experiments demonstrate that our model performs favorably against other state-of-the-art methods, thus confirming the effectiveness of frequency-temporal prior modeling.
****

Unified Visual Relationship Detection with Vision and Language Models
Long Zhao, Liangzhe Yuan, Boqing Gong, Yin Cui, Florian Schroff, Ming-Hsuan Yang, Hartwig Adam, Ting Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6962-6973
This work focuses on training a single visual relationship detector predicting over the union of label spaces from multiple datasets. Merging labels spanning different datasets could be challenging due to inconsistent taxonomies. The issue is exacerbated in visual relationship detection when second-order visual semantics are introduced between pairs of objects. To address this challenge, we propose UniVRD, a novel bottom-up method for Unified Visual Relationship Detection by leveraging vision and language models (VLMs). VLMs provide well-aligned image and text embeddings, where similar relationships are optimized to be close to each other for semantic unification. Our bottom-up design enables the model to enjoy the benefit of training with both object detection and visual relationship datasets. Empirical results on both human-object interaction detection and scene-graph generation demonstrate the competitive performance of our model. UniVRD achieves 38.07 mAP on HICO-DET, outperforming the current best bottom-up HOI detector by 14.26 mAP. More importantly, we show that our unified detector performs as well as dataset-specific models in mAP, and achieves further improvements when we scale up the model. Our code will be made publicly available on GitHub.
****

Occ^2Net: Robust Image Matching Based on 3D Occupancy Estimation for Occluded Regions
Miao Fan, Mingrui Chen, Chen Hu, Shuchang Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9652-9662
Image matching is a fundamental and critical task in various visual applications, such as Simultaneous Localization and Mapping (SLAM) and image retrieval, which require accurate pose estimation. However, most existing methods ignore the occlusion relations between objects caused by camera motion and scene structure.
In this paper, we propose Occ^2Net, a novel image matching method that models occlusion relations using 3D occupancy and infers matching points in occluded regions.
Thanks to the inductive bias encoded in the Occupancy Estimation (OE) module, it greatly simplifies bootstrapping of a multi-view consistent 3D representation

that can then integrate information from multiple views. Together with an Occlus
ion-Aware (OA) module, it incorporates attention layers and rotation alignment t
o enable matching between occluded and visible points.
 We evaluate our method on both real-world and simulated datasets and demonstrat
e its superior performance over state-of-the-art methods on several metrics, esp
ecially in occlusion scenarios.
****
Make-An-Animation: Large-Scale Text-conditional 3D Human Motion Generation
Samaneh Azadi, Akbar Shah, Thomas Hayes, Devi Parikh, Sonal Gupta; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 150
39-15048
Text-guided human motion generation has drawn significant interest because of it
s impactful applications spanning animation and robotics. Recently, application
of diffusion models for motion generation has enabled improvements in the qualit
y of generated motions. However, existing approaches are limited by their relian
ce on relatively small-scale motion capture data, leading to poor performance on
 more diverse, in-the-wild prompts. In this paper, we introduce Make-An-Animatio
n, a text-conditioned human motion generation model which learns more diverse po
ses and prompts from large-scale image-text datasets, enabling significant impro
vement in performance over prior works. Make-An-Animation is trained in two stag
es. First, we train on a curated large-scale dataset of (text, static pseudo-pos
e) pairs extracted from image-text datasets. Second, we fine-tune on motion capt
ure data, adding additional layers to model the temporal dimension. Unlike prior
 diffusion models for motion generation, Make-An-Animation uses a U-Net architec
ture similar to recent text-to-video generation models. Human evaluation of moti
on realism and alignment with input text shows that our model reaches state-of-t
he-art performance on text-to-motion generation.
****
Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image
 Synthesis
Lukas Struppek, Dominik Hintersdorf, Kristian Kersting; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2023, pp. 4584-4596
While text-to-image synthesis currently enjoys great popularity among researcher
s and the general public, the security of these models has been neglected so far
. Many text-guided image generation models rely on pre-trained text encoders fro
m external sources, and their users trust that the retrieved models will behave
as promised. Unfortunately, this might not be the case. We introduce backdoor at
tacks against text-guided generative models and demonstrate that their text enco
ders pose a major tampering risk. Our attacks only slightly alter an encoder so
that no suspicious model behavior is apparent for image generations with clean p
rompts. By then inserting a single character trigger into the prompt, e.g., a no
n-Latin character or emoji, the adversary can trigger the model to either genera
te images with pre-defined attributes or images following a hidden, potentially
malicious description. We empirically demonstrate the high effectiveness of our
attacks on Stable Diffusion and highlight that the injection process of a single
 backdoor takes less than two minutes. Besides phrasing our approach solely as a
n attack, it can also force an encoder to forget phrases related to certain conc
epts, such as nudity or violence, and help to make image generation safer.
****
LD-ZNet: A Latent Diffusion Approach for Text-Based Image Segmentation
Koutilya PNVR, Bharat Singh, Pallabi Ghosh, Behjat Siddiquie, David Jacobs; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023
, pp. 4157-4168
Large-scale pre-training tasks like image classification, captioning, or self-su
pervised techniques do not incentivize learning the semantic boundaries of objec
ts. However, recent generative foundation models built using text-based latent d
iffusion techniques may learn semantic boundaries. This is because they have to
synthesize intricate details about all objects in an image based on a text descr
iption. Therefore, we present a technique for segmenting real and AI-generated i
mages using latent diffusion models (LDMs) trained on internet-scale datasets. F

irst, we show that the latent space of LDMs (z-space) is a better input represen
tation compared to other feature representations like RGB images or CLIP encodin
gs for text-based image segmentation. By training the segmentation models on the
 latent z-space, which creates a compressed representation across several domain
s like different forms of art, cartoons, illustrations, and photographs, we are
also able to bridge the domain gap between real and AI-generated images. We show
 that the internal features of LDMs contain rich semantic information and presen
t a technique in the form of LD-ZNet to further boost the performance of text-ba
sed segmentation. Overall, we show up to 6% improvement over standard baselines
for text-to-image segmentation on natural images. For AI-generated imagery, we s
how close to 20% improvement compared to state-of-the-art techniques. The projec
t is available at https://koutilya-pnvr.github.io/LD-ZNet/.
****
Workie-Talkie: Accelerating Federated Learning by Overlapping Computing and Comm
unications via Contrastive Regularization
Rui Chen, Qiyu Wan, Pavana Prakash, Lan Zhang, Xu Yuan, Yanmin Gong, Xin Fu, Mia
o Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2023, pp. 16999-17009
Federated learning (FL) over mobile devices is a promising distributed learning
paradigm for various mobile applications. However, practical deployment of FL ov
er mobile devices is very challenging because (i) conventional FL incurs huge tr
aining latency for mobile devices due to interleaved local computing and communi
cations of model updates, (ii) there are heterogeneous training data across mobi
le devices, and (iii) mobile devices have hardware heterogeneity in terms of com
puting and communication capabilities. To address aforementioned challenges, in
this paper, we propose a novel "workie-talkie" FL scheme, which can accelerate F
L's training by overlapping local computing and wireless communications via cont
rastive regularization (FedCR). FedCR can reduce FL's training latency and almos
t eliminate straggler issues since it buries/embeds the time consumption of comm
unications into that of local training. To resolve the issue of model staleness
and data heterogeneity co-existing, we introduce class-wise contrastive regulari
zation to correct the local training in FedCR. Besides, we jointly exploit contr
astive regularization and subnetworks to further extend our FedCR approach to ac
commodate edge devices with hardware heterogeneity. We deploy FedCR in our FL te
stbed and conduct extensive experiments. The results show that FedCR outperforms
 its status quo FL approaches on various datasets and models.
****
Downstream-agnostic Adversarial Examples
Ziqi Zhou, Shengshan Hu, Ruizhi Zhao, Qian Wang, Leo Yu Zhang, Junhui Hou, Hai J
in; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC
V), 2023, pp. 4345-4355
Self-supervised learning usually uses a large amount of unlabeled data to pre-tr
ain an encoder which can be used as a general-purpose feature extractor, such th
at downstream users only need to perform fine-tuning operations to enjoy the ben
efit of "big model". Despite this promising prospect, the security of pre-traine
d encoder has not been thoroughly investigated yet, especially when the pre-trai
ned encoder is publicly available for commercial use.
 In this paper, we propose AdvEncoder, the first framework for generating downst
ream-agnostic universal adversarial examples based on the pre-trained encoder. A
dvEncoder aims to construct a universal adversarial perturbation or patch for a
set of natural images that can fool all the downstream tasks inheriting the vict
im pre-trained encoder. Unlike traditional adversarial example works, the pre-tr
ained encoder only outputs feature vectors rather than classification labels. Th
erefore, we first exploit the high frequency component information of the image
to guide the generation of adversarial examples. Then we design a generative att
ack framework to construct adversarial perturbations/patches by learning the dis
tribution of the attack surrogate dataset to improve their attack success rates
and transferability. Our results show that an attacker can successfully attack d
ownstream tasks without knowing either the pre-training dataset or the downstrea
m dataset. We also tailor four defenses for pre-trained encoders, the results of

which further prove the attack ability of AdvEncoder.
****
Late Stopping: Avoiding Confidently Learning from Mislabeled Examples
Suqin Yuan, Lei Feng, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16079-16088
Sample selection is a prevalent method in learning with noisy labels, where small-loss data are typically considered as correctly labeled data. However, this method may not effectively identify clean hard examples with large losses, which are critical for achieving the model's closeto-optimal generalization performance. In this paper, we propose a new framework, Late Stopping, which leverages the intrinsic robust learning ability of DNNs through a prolonged training process. Specifically, Late Stopping gradually shrinks the noisy dataset by removing high-probability mislabeled examples while retaining the majority of clean hard examples in the training set throughout the learning process. We empirically observe that mislabeled and clean examples exhibit differences in the number of epochs required for them to be consistently and correctly classified, and thus high-probability mislabeled examples can be removed. Experimental results on benchmark-simulated and real-world noisy datasets demonstrate that the proposed method outperforms state-of-the-art counterparts.
****
AerialVLN: Vision-and-Language Navigation for UAVs
Shubo Liu, Hongsheng Zhang, Yuankai Qi, Peng Wang, Yanning Zhang, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15384-15394
Recently emerged Vision-and-Language Navigation(VLN) tasks have drawn significant attention in both computer vision and natural language processing communities. Existing VLN tasks are built for agents that navigate on the ground, either indoors or outdoors. However, many tasks require intelligent agents to carry out in the sky, such as UAV-based goods delivery, traffic/security patrol, and scenery tour, to name a few. Navigating in the sky is more complicated than on the ground because agents need to consider the flying height and more complex spatial relationship reasoning. To fill this gap and facilitate research in this field, we propose a new task named AerialVLN, which is UAV-based and towards outdoor environments. We develop a 3D simulator rendered by near-realistic pictures of 25 city-level scenarios. Our simulator supports continuous navigation, environment extension and configuration. We also proposed an extended baseline model based on the widely-used cross modal-alignment (CMA) navigation methods. We find that there is still a significant gap between the baseline model and human performance, which suggests AerialVLN is a new challenging task.
****
On the Robustness of Open-World Test-Time Training: Self-Training with Dynamic Prototype Expansion
Yushu Li, Xun Xu, Yongyi Su, Kui Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11836-11846
Generalizing deep learning models to unknown target domain distribution with low latency has motivated research into test-time training/adaptation (TTT/TTA). Existing approaches often focus on improving test-time training performance under well-curated target domain data. As figured out in this work, many state-of-the-art methods fail to maintain the performance when the target domain is contaminated with strong out-of-distribution (OOD) data, a.k.a. open-world test-time training (OWTTT). The failure is mainly due to the inability to distinguish strong OOD samples from regular weak OOD samples. To improve the robustness of OWTTT we first develop an adaptive strong OOD pruning which improves the efficacy of the self-training TTT method. We further propose a way to dynamically expand the prototypes to represent strong OOD samples for an improved weak/strong OOD data separation. Finally, we regularize self-training with distribution alignment and the combination yields the state-of-the-art performance on 5 OWTTT benchmarks. The code is available at https://github.com/Yushu-Li/OWTTT.
****
Studying How to Efficiently and Effectively Guide Models with Explanations

Sukrut Rao, Moritz Böhle, Amin Parchami-Araghi, Bernt Schiele; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1922-19
33
Despite being highly performant, deep neural networks might base their decisions
 on features that spuriously correlate with the provided labels, thus hurting ge
neralization. To mitigate this, 'model guidance' has recently gained popularity,
 i.e. the idea of regularizing the models' explanations to ensure that they are
"right for the right reasons". While various techniques to achieve such model gu
idance have been proposed, experimental validation of these approaches has thus
far been limited to relatively simple and / or synthetic datasets. To better und
erstand the effectiveness of the various design choices that have been explored
in the context of model guidance, in this work we conduct an in-depth evaluation
 across various loss functions, attribution methods, models, and 'guidance depth
s' on the PASCAL VOC 2007 and MS COCO 2014 datasets. As annotation costs for mod
el guidance can limit its applicability, we also place a particular focus on eff
iciency. Specifically, we guide the models via bounding box annotations, which a
re much cheaper to obtain than the commonly used segmentation masks, and evaluat
e the robustness of model guidance under limited (e.g. with only 1% of annotated
 images) or overly coarse annotations. Further, we propose using the EPG score a
s an additional evaluation metric and loss function ('Energy loss'). We show tha
t optimizing for the Energy loss leads to models that exhibit a distinct focus o
n object-specific features, despite only using bounding box annotations that als
o include background regions. Lastly, we show that such model guidance can impro
ve generalization under distribution shifts. Code available at: https://github.c
om/sukrutrao/Model-Guidance
****
Most Important Person-Guided Dual-Branch Cross-Patch Attention for Group Affect
Recognition
Hongxia Xie, Ming-Xian Lee, Tzu-Jui Chen, Hung-Jen Chen, Hou-I Liu, Hong-Han Shu
ai, Wen-Huang Cheng; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2023, pp. 20598-20608
Group affect refers to the subjective emotion that is evoked by an external stim
ulus in a group, which is an important factor that shapes group behavior and out
comes. Recognizing group affect involves identifying important individuals and s
alient objects among a crowd that can evoke emotions. However, most existing met
hods lack attention to affective meaning in group dynamics and fail to account f
or the contextual relevance of faces and objects in group-level images. In this
work, we propose a solution by incorporating the psychological concept of the Mo
st Important Person (MIP), which represents the most noteworthy face in a crowd
and has affective semantic meaning. We present the Dual-branch Cross-Patch Atten
tion Transformer (DCAT) which uses global image and MIP together as inputs. Spec
ifically, we first learn the informative facial regions produced by the MIP and
the global context separately. Then, the Cross-Patch Attention module is propose
d to fuse the features of MIP and global context together to complement each oth
er. Our proposed method outperforms state-of-the-art methods on GAF 3.0, GroupEm
oW, and HECO datasets. Moreover, we demonstrate the potential for broader applic
ations by showing that our proposed model can be transferred to another group af
fect task, group cohesion, and achieve comparable results.
****
SkeletonMAE: Graph-based Masked Autoencoder for Skeleton Sequence Pre-training
Hong Yan, Yang Liu, Yushen Wei, Zhen Li, Guanbin Li, Liang Lin; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5606-5
618
Skeleton sequence representation learning has shown great advantages for action
recognition due to its promising ability to model human joints and topology. How
ever, the current methods usually require sufficient labeled data for training c
omputationally expensive models. Moreover, these methods ignore how to utilize t
he fine-grained dependencies among different skeleton joints to pre-train an eff
icient skeleton sequence learning model that can generalize well across differen
t datasets. In this paper, we propose an efficient skeleton sequence learning fr

amework, named Skeleton Sequence Learning (SSL). To comprehensively capture the human pose and obtain discriminative skeleton sequence representation, we build an asymmetric graph-based encoder-decoder pre-training architecture named SkeletonMAE, which embeds skeleton joint sequence into graph convolutional network and reconstructs the masked skeleton joints and edges based on the prior human topology knowledge. Then, the pre-trained SkeletonMAE encoder is integrated with the Spatial-Temporal Representation Learning (STRL) module to build the SSL framework. Extensive experimental results show that our SSL generalizes well across different datasets and outperforms the state-of-the-art self-supervised skeleton-based methods on FineGym, Diving48, NTU 60 and NTU 120 datasets. Moreover, we obtain comparable performance to some fully supervised methods. The code is avaliable at https://github.com/HongYan1123/SkeletonMAE.
****
Achievement-Based Training Progress Balancing for Multi-Task Learning
Hayoung Yun, Hanjoo Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16935-16944
Multi-task learning faces two challenging issues: (1) the high cost of annotating labels for all tasks and (2) balancing the training progress of various tasks with different natures. To resolve the label annotation issue, we construct a large-scale "partially annotated" multi-task dataset by combining task-specific datasets. However, the numbers of annotations for individual tasks are imbalanced, which
may escalate an imbalance in training progress. To balance the training progress, we propose an achievement-based multi-task loss to modulate training speed based on the "achievement," defined as the ratio of current accuracy to single-task accuracy. Then, we formulate the multitask loss as a weighted geometric mean of individual task losses instead of a weighted sum to prevent any task from dominating the loss. In experiments, we evaluated the accuracy and training speed of the proposed multi-task loss on the large-scale multi-task dataset against recent multitask losses. The proposed loss achieved the best multi-task accuracy without incurring training time overhead. Compared to single-task models, the proposed one achieved 1.28%, 1.65%, and 1.18% accuracy improvement in object detection, semantic segmentation, and depth estimation, respectively, while reducing computations to 33.73%. Source code is available at https://github.com/samsung/Achievement-based-MTL.
****
Pose-Free Neural Radiance Fields via Implicit Pose Regularization
Jiahui Zhang, Fangneng Zhan, Yingchen Yu, Kunhao Liu, Rongliang Wu, Xiaoqin Zhang, Ling Shao, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3534-3543
Pose-free neural radiance fields (NeRF) aim to train NeRF with unposed multi-view images and it has achieved very impressive success in recent years. Most existing works share the pipeline of training a coarse pose estimator with rendered images at first, followed by a joint optimization of estimated poses and neural radiance field. However, as the pose estimator is trained with only rendered images, the pose estimation is usually biased or inaccurate for real images due to the domain gap between real images and rendered images, leading to poor robustness for the pose estimation of real images and further local min- ima in joint optimization. We design IR-NeRF, an innovative pose-free NeRF that introduces implicit pose regularization to refine pose estimator with unposed real images and improve the robustness of the pose estimation for real images. With a collection of 2D images of a specific scene, IR-NeRF constructs a scene codebook that stores scene features and captures the scene-specific pose distribution implicitly as priors. Thus, the robustness of pose estimation can be promoted with the scene priors according to the rationale that a 2D real image can be well reconstructed from the scene codebook only when its estimated pose lies within the pose distribution. Extensive experiments show that IR-NeRF achieves superior novel view synthesis and outperforms the state-of-the-art consistently across multiple synthetic and real datasets.
****

Self-supervised Learning to Bring Dual Reversed Rolling Shutter Images Alive
Wei Shang, Dongwei Ren, Chaoyu Feng, Xiaotao Wang, Lei Lei, Wangmeng Zuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13086-13094
Modern consumer cameras usually employ the rolling shutter (RS) mechanism, where images are captured by scanning scenes row-by-row, yielding RS distortions for dynamic scenes. To correct RS distortions, existing methods adopt a fully supervised learning manner, where high framerate global shutter (GS) images should be collected as ground-truth supervision. In this paper, we propose a Self-supervised learning framework for Dual reversed RS distortions Correction (SelfDRSC), where a DRSC network can be learned to generate a high framerate GS video only based on dual RS images with reversed distortions. In particular, a bidirectional distortion warping module is proposed for reconstructing dual reversed RS images, and then a self-supervised loss can be deployed to train DRSC network by enhancing the cycle consistency between input and reconstructed dual reversed RS images. Besides start and end RS scanning time, GS images at arbitrary intermediate scanning time can also be supervised in SelfDRSC, thus enabling the learned DRSC network to generate a high framerate GS video. Moreover, a simple yet effective self-distillation strategy is introduced in self-supervised loss for mitigating boundary artifacts in generated GS images. On synthetic dataset, SelfDRSC achieves better or comparable quantitative metrics in comparison to state-of-the-art methods trained in the full supervision manner. On real-world RS cases, our SelfDRSC can produce high framerate GS videos with finer correction textures and better temporary consistency. The source code and trained models are made publicly available at https://github.com/ shangwei5/SelfDRSC.
****
Logic-induced Diagnostic Reasoning for Semi-supervised Semantic Segmentation
Chen Liang, Wenguan Wang, Jiaxu Miao, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16197-16208
Recent advances in semi-supervised semantic segmentation have been heavily reliant on pseudo labeling to compensate for limited labeled data, disregarding the valuable relational knowledge among semantic concepts. To bridge this gap, we devise LogicDiag, a brand new neural-logic semi-supervised learning framework. Our key insight is that conflicts within pseudo labels, identified through symbolic knowledge, can serve as strong yet commonly ignored learning signals. LogicDiag resolves such conflicts via reasoning with logic-induced diagnoses, enabling the recovery of (potentially) erroneous pseudo labels, ultimately alleviating the notorious error accumulation problem. We showcase the practical application of LogicDiag in the data-hungry segmentation scenario, where we formalize the structured abstraction of semantic concepts as a set of logic rules. Extensive experiments on three standard semi-supervised semantic segmentation benchmarks demonstrate the effectiveness and generality of LogicDiag. Moreover, LogicDiag highlights the promising opportunities arising from the systematic integration of symbolic reasoning into the prevalent statistical, neural learning approaches.
****
Self-Supervised Monocular Depth Estimation by Direction-aware Cumulative Convolution Network
Wencheng Han, Junbo Yin, Jianbing Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8613-8623
Monocular depth estimation is known as an ill-posed task that objects in a 2D image usually do not contain sufficient information to predict their depth. Thus, it acts differently from other tasks (e.g., classification and segmentation) in many ways. In this paper, we find that self-supervised monocular depth estimation shows a direction sensitivity and environmental dependency in the feature representation. But the current CNN backbones borrowed from other tasks cannot handle different types of environmental information efficiently, limiting the overall depth accuracy. To bridge this gap, we propose a new Direction-aware Cumulative Convolution Network (DaCCN), which improves the depth feature representation in two aspects. First, we propose a direction-aware module, which can learn to adjust the feature extraction in each direction, facilitating the encoding of diffe

rent types of information. Secondly, we design a new cumulative convolution to improve the efficiency for aggregating important environmental information. Experiments show that our method achieves significant improvements on three widely used benchmarks and sets a new state-of-the-art performance on the popular benchmarks with all three types of self-supervision.
****
Encyclopedic VQA: Visual Questions About Detailed Properties of Fine-Grained Categories
Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, Vittorio Ferrari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3113-3124
We propose Encyclopedic-VQA, a large scale visual question answering (VQA) dataset featuring visual questions about detailed properties of fine-grained categories and instances. It contains 221k unique question+answer pairs each matched with (up to) 5 images, resulting in a total of 1M VQA samples. Moreover, our dataset comes with a controlled knowledge base derived from Wikipedia, marking the evidence to support each answer. Empirically, we show that our dataset poses a hard challenge for large vision+language models as they perform poorly on our dataset: PaLI [9] is state-of-the-art on OK-VQA [29], yet it only achieves 13.0% accuracy on our dataset. Moreover, we experimentally show that progress on answering our encyclopedic questions can be achieved by augmenting large models with a mechanism that retrieves relevant information for the knowledge base. An oracle experiment with perfect retrieval achieves 87.0% accuracy on the single-hop portion of our dataset, and an automatic retrieval-
augmented prototype yields 48.8%. We believe that our dataset enables future research on retrieval-augmented vision+language models.
****
Towards Understanding the Generalization of Deepfake Detectors from a Game-Theoretical View
Kelu Yao, Jin Wang, Boyu Diao, Chao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2031-2041
This paper aims to explain the generalization of deepfake detectors from the novel perspective of multi-order interactions among visual concepts. Specifically, we propose three hypotheses:
1. Deepfake detectors encode multi-order interactions among visual concepts, in which the low-order interactions usually have substantially negative contributions to deepfake detection.
2. Deepfake detectors with better generalization abilities tend to encode low-order interactions with fewer negative contributions.
3. Generalized deepfake detectors usually weaken the negative contributions of low-order interactions by suppressing their strength.

Accordingly, we design several mathematical metrics to evaluate the effect of low-order interaction for deepfake detectors.
Extensive comparative experiments are conducted, which verify the soundness of our hypotheses.
Based on the analyses, we further propose a generic method, which directly reduces the toxic effects of low-order interactions to improve the generalization of deepfake detectors to some extent.
The code will be released when the paper is accepted.
****
Few-Shot Common Action Localization via Cross-Attentional Fusion of Context and Temporal Dynamics
Juntae Lee, Mihir Jain, Sungrack Yun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10214-10223
The goal of this paper is to localize action instances in a long untrimmed query video using just meager trimmed support videos representing a common action whose class information is not given. In this task, it is crucial to mine reliable temporal cues representing a common action from handful support videos. In our work, we develop an attention mechanism using cross-correlation. Based on this cross-attention, we first transform the support videos into query video's context

to emphasize query-relevant important frames, and suppress less relevant ones. N
ext, we summarize sub-sequences of support video frames to represent temporal dy
namics in coarse temporal granularity, which is then propagated to the fine-grai
ned support video features through the cross-attention. In each case, the cross-
attentions are applied to each support video in the individual-to-all strategy t
o balance heterogeneity and compatibility of the support videos. In contrast, th
e candidate instances in the query video are lastly attended by the resulting su
pport video features, at once. In addition, we also develop a relational classif
ier head based on the query and support video representations. We show the effec
tiveness of our work with the state-of-the-art (SOTA) performance in benchmark d
atasets (ActivityNet1.3 and THUMOS14), and analyze each component extensively.
****
Physically-Plausible Illumination Distribution Estimation
Egor Ershov, Vasily Tesalin, Ivan Ermakov, Michael S. Brown; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12928-129
36
A camera's auto-white-balance (AWB) module operates under the assumption that th
ere is a single dominant illumination in a captured scene. AWB methods estimate
an image's dominant illumination and use it as the target "white point" for corr
ection. However, in natural scenes, there are often many light sources present.
We performed a user study that revealed that non-dominant illuminations often pr
oduce visually pleasing white-balanced images and, in some cases, are even prefe
rred over the dominant illumination. Motivated by this observation, we revisit A
WB to predict a distribution of plausible illuminations for use in white balan
ce. As part of this effort, we extend the Cube++ illumination estimation dataset
 to provide ground truth illumination distributions per image. Using this new gr
ound truth data, we describe how to train a lightweight neural network method to
 predict the scene's illumination distribution. We describe how our idea can be
used with existing image formats by embedding the estimated distribution in the
RAW image to enable users to generate visually plausible white-balance images.
****
3DPPE: 3D Point Positional Encoding for Transformer-based Multi-Camera 3D Object
 Detection
Changyong Shu, Jiajun Deng, Fisher Yu, Yifan Liu; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2023, pp. 3580-3589
Transformer-based methods have swept the benchmarks on 2D and 3D detection on im
ages. Because tokenization before the attention mechanism drops the spatial info
rmation, positional encoding becomes critical for those methods. Recent works fo
und that encodings based on samples of the 3D viewing rays can significantly imp
rove the quality of multi-camera 3D object detection. We hypothesize that 3D poi
nt locations can provide more information than rays. Therefore, we introduce 3D
point positional encoding, 3DPPE, to the 3D detection Transformer decoder. Altho
ugh 3D measurements are not available at the inference time of monocular 3D obje
ct detection, 3DPPE uses predicted depth to approximate the real point positions
. Our hybrid-depth module combines direct and categorical depth to estimate the
refined depth of each pixel. Despite the approximation, 3DPPE achieves 46.0 mAP
and 51.4 NDS on the competitive nuScenes dataset, significantly outperforming en
codings based on ray samples. The codes are available at https://github.com/dril
istbox/3DPPE.
****
Revisiting Foreground and Background Separation in Weakly-supervised Temporal Ac
tion Localization: A Clustering-based Approach
Qinying Liu, Zilei Wang, Shenghai Rong, Junjie Li, Yixin Zhang; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10433-
10443
Weakly-supervised temporal action localization aims to localize action instances
 in videos with only video-level action labels. Existing methods mainly embrace
a localization-by-classification pipeline that optimizes the snippet-level predi
ction with a video classification loss. However, this formulation suffers from t
he discrepancy between classification and detection, resulting in inaccurate sep

aration of foreground and background (F&B) snippets. To alleviate this problem, we propose to explore the underlying structure among the snippets by resorting to unsupervised snippet clustering, rather than heavily relying on the video classification loss. Specifically, we propose a novel clustering-based F&B separation algorithm. It comprises two core components: a snippet clustering component that groups the snippets into multiple latent clusters and a cluster classification component that further classifies the cluster as foreground or background. As there are no ground-truth labels to train these two components, we introduce a unified self-labeling mechanism based on optimal transport to produce high-quality pseudo-labels that match several plausible prior distributions. This ensures that the cluster assignments of the snippets can be accurately associated with their F&B labels, thereby boosting the F&B separation. We evaluate our method on three benchmarks: THUMOS14, ActivityNet v1.2 and v1.3. Our method achieves promising performance on all three benchmarks while being significantly more lightweight than previous methods. Code is available at https://github.com/Qinying-Liu/CASE
****
VertexSerum: Poisoning Graph Neural Networks for Link Inference
Ruyi Ding, Shijin Duan, Xiaolin Xu, Yunsi Fei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4532-4541
Graph neural networks (GNNs) have brought superb performance to various applications utilizing graph structural data, such as social analysis and fraud detection. The graph links, e.g., social relationships and transaction history, are sensitive and valuable information, which raises privacy concerns when using GNNs. To exploit these vulnerabilities, we propose VertexSerum, a novel graph poisoning attack that increases the effectiveness of graph link stealing by amplifying the link connectivity leakage. To infer node adjacency more accurately, we propose an attention mechanism that can be embedded into the link detection network. Our experiments demonstrate that VertexSerum significantly outperforms the SOTA link inference attack, improving the AUC scores by an average of 9.8% across four real-world datasets and three different GNN structures. Furthermore, our experiments reveal the effectiveness of VertexSerum in both black-box and online learning settings, further validating its applicability in real-world scenarios.
****
NeRF-Det: Learning Geometry-Aware Volumetric Representation for Multi-View 3D Object Detection
Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, Masayoshi Tomizuka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23320-23330
We present NeRF-Det, a novel method for indoor 3D detection with posed RGB images as input. Unlike existing indoor 3D detection methods that struggle to model scene geometry, our method makes novel use of NeRF in an end-to-end manner to explicitly estimate 3D geometry, thereby improving 3D detection performance. Specifically, to avoid the significant extra latency associated with per-scene optimization of NeRF, we introduce sufficient geometry priors to enhance the generalizability of NeRF-MLP. Furthermore, we subtly connect the detection and NeRF branches through a shared MLP, enabling an efficient adaptation of NeRF to detection and yielding geometry-aware volumetric representations for 3D detection. Our method outperforms state-of-the-arts by 3.9 mAP and 3.1 mAP on the ScanNet and ARKITScenes benchmarks, respectively. We provide extensive analysis to shed light on how NeRF-Det works. As a result of our joint-training design, NeRF-Det is able to generalize well to unseen scenes for object detection, view synthesis, and depth estimation tasks without requiring per-scene optimization. Code is available at https://github.com/facebookresearch/NeRF-Det.
****
Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception
Kun Yang, Dingkang Yang, Jingyu Zhang, Mingcheng Li, Yang Liu, Jing Liu, Hanqi Wang, Peng Sun, Liang Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23383-23392
Multi-agent collaborative perception as a potential application for vehicle-to-e

verything communication could significantly improve the perception performance of autonomous vehicles over single-agent perception. However, several challenges remain in achieving pragmatic information sharing in this emerging research. In this paper, we propose SCOPE, a novel collaborative perception framework that aggregates the spatio-temporal awareness characteristics across on-road agents in an end-to-end manner. Specifically, SCOPE has three distinct strengths: i) it considers effective semantic cues of the temporal context to enhance current representations of the target agent; ii) it aggregates perceptually critical spatial information from heterogeneous agents and overcomes localization errors via multi-scale feature interactions; iii) it integrates multi-source representations of the target agent based on their complementary contributions by an adaptive fusion paradigm. To thoroughly evaluate SCOPE, we consider both real-world and simulated scenarios of collaborative 3D object detection tasks on three datasets. Extensive experiments show the superiority of our approach and the necessity of the proposed components. The project link is https://ydk122024.github.io/SCOPE/.
****

LPFF: A Portrait Dataset for Face Generators Across Large Poses
Yiqian Wu, Jing Zhang, Hongbo Fu, Xiaogang Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20327-20337
Existing face generators exhibit exceptional performance on faces in small to medium poses (with respect to frontal faces) but struggle to produce realistic results for large poses. The distorted rendering results on large poses in 3D-aware generators further show that the generated 3D face shapes are far from the distribution of 3D faces in reality. We find that the above issues are caused by the training dataset's pose imbalance. To this end, we present LPFF, a large-pose Flickr face dataset comprised of 19,590 high-quality real large-pose portrait images. We utilize our dataset to train a 2D face generator that can process large-pose face images, as well as a 3D-aware generator that can generate realistic human face geometry. To better validate our pose-conditional 3D-aware generators, we develop a new FID measure to evaluate the 3D-level performance. Through this novel FID measure and other experiments, we show that LPFF can help 2D face generators extend their latent space and better manipulate the large-pose data, and help 3D-aware face generators achieve better view consistency and more realistic 3D reconstruction results.
****

Pseudo-label Alignment for Semi-supervised Instance Segmentation
Jie Hu, Chen Chen, Liujuan Cao, Shengchuan Zhang, Annan Shu, Guannan Jiang, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16337-16347
Pseudo-labeling is significant for semi-supervised instance segmentation, which generates instance masks and classes from unannotated images for subsequent training. However, in existing pipelines, pseudo-labels that contain valuable information may be directly filtered out due to mismatches in class and mask quality. To address this issue, we propose a novel framework, called pseudo-label aligning instance segmentation (PAIS), in this paper. In PAIS, we devise a dynamic aligning loss (DALoss) that adjusts the weights of semi-supervised loss terms with varying class and mask score pairs. Through extensive experiments conducted on the COCO and Cityscapes datasets, we demonstrate that PAIS is a promising framework for semi-supervised instance segmentation, particularly in cases where labeled data is severely limited. Notably, with just 1% labeled data, PAIS achieves 21.2 mAP (based on Mask-RCNN) and 19.9 mAP (based on K-Net) on the COCO dataset, outperforming the current state-of-the-art model, i.e., NoisyBoundary with 7.7 mAP, by a margin of over 12 points. Code is available at: https://github.com/hujiecpp/PAIS.
****

Deep Geometrized Cartoon Line Inbetweening
Li Siyao, Tianpei Gu, Weiye Xiao, Henghui Ding, Ziwei Liu, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7291-7300
We aim to address a significant but understudied problem in the anime industry,

namely the inbetweening of cartoon line drawings. Inbetweening involves generati
ng intermediate frames between two black-and-white line drawings and is a time-c
onsuming and expensive process that can benefit from automation. However, existi
ng frame interpolation methods that rely on matching and warping whole raster im
ages are unsuitable for line inbetweening and often produce blurring artifacts t
hat damage the intricate line structures. To preserve the precision and detail o
f the line drawings, we propose a new approach, called AnimeInbet, which geometr
izes raster line drawings into graphs of endpoints and reframes the inbetweening
 task as a graph fusion problem with vertex repositioning. Our method can effect
ively capture the sparsity and unique structure of line drawings while preservin
g the details during inbetweening. This is made possible through our novel modul
es, i.e., vertex encoding, a vertex correspondence Transformer, an effective mec
hanism for vertex repositioning and a visibility predictor. To train our method,
 we introduce MixamoLine240, a new dataset of line drawings with ground truth ve
ctorization and matching labels. Our experiments demonstrate that AnimeInbet syn
thesizes high-quality, clean, and complete intermediate line drawings, outperfor
ming existing methods quantitatively and qualitatively, especially in cases with
 large motions.
****
MixBag: Bag-Level Data Augmentation for Learning from Label Proportions
Takanori Asanomi, Shinnosuke Matsuo, Daiki Suehiro, Ryoma Bise; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16570-
16579
Learning from label proportions (LLP) is a promising weakly supervised learning
problem. In LLP, a set of instances (bag) has label proportions but no instance-
level labels.
 LLP aims to train an instance-level classifier by using the label proportions o
f the bag.
 In this paper, we propose a bag-level data augmentation method for LLP called M
ixBag, which is based on the key observation from our preliminary experiments; t
hat the instance-level classification accuracy improves as the number of labeled
 bags increases even though the total number of instances is fixed.
 We also propose a confidence interval loss designed based on statistical theory
 in order to use the augmented bags effectively.
 To the best of our knowledge, this is the first attempt to propose bag-level da
ta augmentation for LLP.
 The advantage of MixBag is that it can be applied to instance-level data augmen
tation techniques and any LLP method that uses the proportion loss.
 Experimental results demonstrate this advantage and the effectiveness of our me
thod.
****
Effective Real Image Editing with Accelerated Iterative Diffusion Inversion
Zhihong Pan, Riccardo Gherardi, Xiufeng Xie, Stephen Huang; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15912-1592
1
Despite all recent progress, it is still challenging to edit and manipulate natu
ral images with modern generative models. When using Generative Adversarial Netw
ork (GAN), one major hurdle is in the inversion process mapping a real image to
its corresponding noise vector in the latent space, since its necessary to be ab
le to reconstruct an image to edit its contents. Likewise for Denoising Diffusio
n Implicit Models (DDIM), the linearization assumption in each inversion step ma
kes the whole deterministic inversion process unreliable. Existing approaches th
at have tackled the problem of inversion stability often incur in significant tr
ade-offs in computational efficiency. In this work we propose an Accelerated Ite
rative Diffusion Inversion method, dubbed AIDI, that significantly improves reco
nstruction accuracy with minimal additional overhead in space and time complexit
y. By using a novel blended guidance technique, we show that effective results c
an be obtained on a large range of image editing tasks without large classifier-
free guidance in inversion. Furthermore, when compared with other diffusion inve
rsion based works, our proposed process is shown to be more robust for fast imag

e editing in the 10 and 20 diffusion steps' regimes.
****
3D-Aware Neural Body Fitting for Occlusion Robust 3D Human Pose Estimation
Yi Zhang, Pengliang Ji, Angtian Wang, Jieru Mei, Adam Kortylewski, Alan Yuille; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9399-9410
Regression-based methods for 3D human pose estimation directly predict the 3D pose parameters from a 2D image using deep networks. While achieving state-of-the-art performance on standard benchmarks, their performance degrades under occlusion. In contrast, optimization-based methods fit a parametric body model to 2D features in an iterative manner. The localized reconstruction loss can potentially make them robust to occlusion, but they suffer from the 2D-3D ambiguity. Motivated by the recent success of generative models in rigid object pose estimation, we propose 3D-aware Neural Body Fitting (3DNBF) - an approximate analysis-by-synthesis approach to 3D human pose estimation with SOTA performance and occlusion robustness. In particular, we propose a generative model of deep features based on a volumetric human representation with Gaussian ellipsoidal kernels emitting 3D pose-dependent feature vectors. The neural features are trained with contrastive learning to become 3D-aware and hence to overcome the 2D-3D ambiguity. Experiments show that 3DNBF outperforms other approaches on both occluded and standard benchmarks.
****
Chinese Text Recognition with A Pre-Trained CLIP-Like Model Through Image-IDS Aligning
Haiyang Yu, Xiaocong Wang, Bin Li, Xiangyang Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11943-11952
Scene text recognition has been studied for decades due to its broad applications. However, despite Chinese characters possessing different characteristics from Latin characters, such as complex inner structures and large categories, few methods have been proposed for Chinese Text Recognition (CTR). Particularly, the characteristic of large categories poses challenges in dealing with zero-shot and few-shot Chinese characters. In this paper, inspired by the way humans recognize Chinese texts, we propose a two-stage framework for CTR. Firstly, we pre-train a CLIP-like model through aligning printed character images and Ideographic Description Sequences (IDS). This pre-training stage simulates humans recognizing Chinese characters and obtains the canonical representation of each character. Subsequently, the learned representations are employed to supervise the CTR model, such that traditional single-character recognition can be improved to text-line recognition through image-IDS matching. To evaluate the effectiveness of the proposed method, we conduct extensive experiments on both Chinese character recognition (CCR) and CTR. The experimental results demonstrate that the proposed method performs best in CCR and outperforms previous methods in most scenarios of the CTR benchmark. It is worth noting that the proposed method can recognize zero-shot Chinese characters in text images without fine-tuning, whereas previous methods require fine-tuning when new classes appear. The code is available at https://github.com/FudanVI/FudanOCR/tree/main/image-ids-CTR.
****
MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond
Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, Bo Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3205-3215
Neural radiance fields (NeRF) and its subsequent variants have led to remarkable progress in neural rendering. While most of recent neural rendering works focus on objects and small-scale scenes, developing neural rendering methods for city-scale scenes is of great potential in many real-world applications. However, this line of research is impeded by the absence of a comprehensive and high-quality dataset, yet collecting such a dataset over real city-scale scenes is costly, sensitive, and technically infeasible. To this end, we build a large-scale, comprehensive, and high-quality synthetic dataset for city-scale neural rendering re

searches. Leveraging the Unreal Engine 5 City Sample project, we developed a pip
eline to easily collect aerial and street city views, accompanied by ground-trut
h camera poses and a range of additional data modalities. Flexible controls on e
nvironmental factors like light, weather, human and car crowd are also available
 in our pipeline, supporting the need of various tasks covering city-scale neura
l rendering and beyond. The resulting pilot dataset, MatrixCity, contains 67k ae
rial images and 452k street images from two city maps of total size 28km^2. On t
op of MatrixCity, a thorough benchmark is also conducted, which not only reveals
 unique challenges of the task of city-scale neural rendering, but also highligh
ts potential improvements for future works. The dataset and code will be publicl
y available at the project page: https://city-super.github.io/matrixcity/.
****
LinkGAN: Linking GAN Latents to Pixels for Controllable Image Synthesis
Jiapeng Zhu, Ceyuan Yang, Yujun Shen, Zifan Shi, Bo Dai, Deli Zhao, Qifeng Chen;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2023, pp. 7656-7666
This work presents an easy-to-use regularizer for GAN training, which helps expl
icitly link some axes of the latent space to a set of pixels in the synthesized
image. Establishing such a connection facilitates a more convenient local contro
l of GAN generation, where users can alter the image content only within a spati
al area simply by partially resampling the latent code. Experimental results con
firm four appealing properties of our regularizer, which we call LinkGAN. (1) Th
e latent-pixel linkage is applicable to either a fixed region (i.e., same for al
l instances) or a particular semantic category (i.e., varying across instances),
 like the sky. (2) Two or multiple regions can be independently linked to differ
ent latent axes, which further supports joint control. (3) Our regularizer can i
mprove the spatial controllability of both 2D and 3D-aware GAN models, barely sa
crificing the synthesis performance. (4) The models trained with our regularizer
 are compatible with GAN inversion techniques and maintain editability on real i
mages.
****
Exploiting Proximity-Aware Tasks for Embodied Social Navigation
Enrico Cancelli, Tommaso Campari, Luciano Serafini, Angel X. Chang, Lamberto Bal
lan; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 10957-10967
Learning how to navigate among humans in an occluded and spatially constrained i
ndoor environment, is a key ability required to embodied agents to be integrated
 into our society. In this paper, we propose an end-to-end architecture that exp
loits Proximity-Aware Tasks (referred as to Risk and Proximity Compass) to injec
t into a reinforcement learning navigation policy the ability to infer common-se
nse social behaviours. To this end, our tasks exploit the notion of immediate an
d future dangers of collision. Furthermore, we propose an evaluation protocol sp
ecifically designed for the Social Navigation Task in simulated environments. Th
is is done to capture fine-grained features and characteristics of the policy by
 analyzing the minimal unit of human-robot spatial interaction, called Encounter
. We validate our approach on Gibson4+ and Habitat-Matterport3D datasets.
****
SVDiff: Compact Parameter Space for Diffusion Fine-Tuning
Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, Feng Yang;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2023, pp. 7323-7334
Recently, diffusion models have achieved remarkable success in text-to-image gen
eration, enabling the creation of high-quality images from text prompts and vari
ous conditions. However, existing methods for customizing these models are limit
ed by handling multiple personalized subjects and the risk of overfitting. Moreo
ver, the large parameter space is inefficient for model storage. In this paper,
we propose a novel approach to address the limitations in existing text-to-image
 diffusion models for personalization and customization. Our method involves fin
e-tuning the singular values of the weight matrices, leading to a compact and ef
ficient parameter space that reduces the risk of overfitting and language-drifti

ng. Our approach also includes a Cut-Mix-Unmix data-augmentation technique to enhance the quality of multi-subject image generation and a simple text-based image editing framework. Our proposed SVDiff method has a significantly smaller model size (1.7MB for StableDiffusion) compared to existing methods, making it more practical for real-world applications.
****
UniFace: Unified Cross-Entropy Loss for Deep Face Recognition
Jiancan Zhou, Xi Jia, Qiufu Li, Linlin Shen, Jinming Duan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20730-20739
As a widely used loss function in deep face recognition, the softmax loss cannot guarantee that the minimum positive sample-to-class similarity is larger than the maximum negative sample-to-class similarity. As a result, no unified threshold is available to separate positive sample-to-class pairs from negative sample-to-class pairs. To bridge this gap, we design a UCE (Unified Cross-Entropy) loss for face recognition model training, which is built on the vital constraint that all the positive sample-to-class similarities shall be larger than the negative ones.
Our UCE loss can be integrated with margins for a further performance boost. The face recognition model trained with the proposed UCE loss, UniFace, was intensively evaluated using a number of popular public datasets like MFR, IJB-C, LFW, CFP-FP, AgeDB, and MegaFace. Experimental results show that our approach outperforms SOTA methods like SphereFace, CosFace, ArcFace, Partial FC, etc. Especially, till the submission of this work (Mar. 8, 2023), the proposed UniFace achieves the highest TAR@MR-All on the academic track of the MFR-ongoing challenge. Code is publicly available.
****
Jumping through Local Minima: Quantization in the Loss Landscape of Vision Transformers
Natalia Frumkin, Dibakar Gope, Diana Marculescu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16978-16988
Quantization scale and bit-width are the most important parameters when considering how to quantize a neural network. Prior work focuses on optimizing quantization scales in a global manner through gradient methods (gradient descent & Hessian analysis). Yet, when applying perturbations to quantization scales, we observe a very jagged, highly non-smooth test loss landscape. In fact, small perturbations in quantization scale can greatly affect accuracy, yielding a 0.5-0.8% accuracy boost in 4-bit quantized vision transformers (ViTs). In this regime, gradient methods break down, since they cannot reliably reach local minima.
In our work, dubbed Evol-Q, we use evolutionary search to effectively traverse the non-smooth landscape.
Additionally, we propose using an infoNCE loss, which not only helps combat overfitting on the small (1,000 images) calibration dataset but also makes traversing such a highly non-smooth surface easier. Evol-Q improves the top-1 accuracy of a fully quantized ViT-Base by 10.30%, 0.78%, and 0.15% for 3-bit, 4-bit, and 8-bit weight quantization levels. Extensive experiments on a variety of CNN and ViT architectures further demonstrate its robustness in extreme quantization scenarios.
****
Hierarchical Contrastive Learning for Pattern-Generalizable Image Corruption Detection
Xin Feng, Yifeng Xu, Guangming Lu, Wenjie Pei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12076-12085
Effective image restoration with large-size corruptions, such as blind image inpainting, entails precise detection of corruption region masks which remains extremely challenging due to diverse shapes and patterns of corruptions. In this work, we present a novel method for automatic corruption detection, which allows for blind corruption restoration without known corruption masks. Specifically, we develop a hierarchical contrastive learning framework to detect corrupted regions by capturing the intrinsic semantic distinctions between corrupted and uncorrupted regions. In particular, our model detects the corrupted mask in a coarse-to

-fine manner by first predicting a coarse mask by contrastive learning in low-resolution feature space and then refines the uncertain area of the mask by high-resolution contrastive learning. A specialized hierarchical interaction mechanism is designed to facilitate the knowledge propagation of contrastive learning in different scales, boosting the modeling performance substantially. The detected multi-scale corruption masks are then leveraged to guide the corruption restoration. Detecting corrupted regions by learning the contrastive distinctions rather than the semantic patterns of corruptions, our model has well generalization ability across different corruption patterns. Extensive experiments demonstrate following merits of our model: 1) the superior performance over other methods on both corruption detection and various image restoration tasks including blind inpainting and watermark removal, and 2) strong generalization across different corruption patterns such as graffiti, random noise or other image content. Codes and trained weights are available at https://github.com/xyfJASON/HCL.

****

Learning Optical Flow from Event Camera with Rendered Dataset

Xinglong Luo, Kunming Luo, Ao Luo, Zhengning Wang, Ping Tan, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9847-9857

We study the problem of estimating optical flow from event cameras. One important issue is how to build a high-quality event-flow dataset with accurate event values and flow labels. Previous datasets are created by either capturing real scenes by event cameras or synthesizing from images with pasted foreground objects. The former case can produce real event values but with calculated flow labels, which are sparse and inaccurate. The later case can generate dense flow labels but the interpolated events are prone to errors. In this work, we propose to render a physically correct event-flow dataset using computer graphics models. In particular, we first create indoor and outdoor 3D scenes by Blender with rich scene content variations. Second, diverse camera motions are included for the virtual capturing, producing images and accurate flow labels. Third, we render high-framerate videos between images for accurate events. The rendered dataset can adjust the density of events, based on which we further introduce an adaptive density module (ADM). Experiments show that our proposed dataset can facilitate event-flow learning, whereas previous approaches when trained on our dataset can improve their performances constantly by a relatively large margin. In addition, event-flow pipelines when equipped with our ADM can further improve performances. Our code and dataset will be publicly available.

****

EPiC: Ensemble of Partial Point Clouds for Robust Classification

Meir Yossef Levi, Guy Gilboa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14475-14484

Robust point cloud classification is crucial for real-world applications,as consumer-type 3D sensors often yield partial and noisy data, degraded by various artifacts. In this work we propose a general ensemble framework, based on partial point cloud sampling. Each ensemble member is exposed to only partial input data. Three sampling strategies are used jointly, two local ones, based on patches and curves, and a global one of random sampling. We demonstrate the robustness of our method to various local and global degradations. We show that our framework significantly improves the robustness of top classification netowrks by a large margin. Our experimental setting uses the recently introduced ModelNet-C database by Ren et al., where we reach SOTA both on unaugmented and on augmented data. Our unaugmented mean Corruption Error (mCE) is 0.64 (current SOTA is 0.86) and 0.50 for augmented data (current SOTA is 0.57). We analyze and explain these remarkable results through diversity analysis. Our code is availabe at: https://github.com/yossilevii100/EPiC

****

Distilling Large Vision-Language Model with Out-of-Distribution Generalizability

Xuanlin Li, Yunhao Fang, Minghua Liu, Zhan Ling, Zhuowen Tu, Hao Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2492-2503

Large vision-language models have achieved outstanding performance, but their size and computational requirements make their deployment on resource-constrained devices and time-sensitive tasks impractical. Model distillation, the process of creating smaller, faster models that maintain the performance of larger models, is a promising direction towards the solution. This paper investigates the distillation of visual representations in large teacher vision-language models into lightweight student models using a small- or mid-scale dataset. Notably, this study focuses on open-vocabulary out-of-distribution (OOD) generalization, a challenging problem that has been overlooked in previous model distillation literature. We propose two principles from vision and language modality perspectives to enhance student's OOD generalization: (1) by better imitating teacher's visual representation space, and carefully promoting better coherence in vision-language alignment with the teacher; (2) by enriching the teacher's language representations with informative and finegrained semantic attributes to effectively distinguish between different labels. We propose several metrics and conduct extensive experiments to investigate their techniques. The results demonstrate significant improvements in zero-shot and few-shot student performance on open-vocabulary out-of-distribution classification, highlighting the effectiveness of our proposed approaches.
****
Cross-Modal Learning with 3D Deformable Attention for Action Recognition
Sangwon Kim, Dasom Ahn, Byoung Chul Ko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10265-10275
An important challenge in vision-based action recognition is the embedding of spatiotemporal features with two or more heterogeneous modalities into a single feature. In this study, we propose a new 3D deformable transformer for action recognition with adaptive spatiotemporal receptive fields and a cross-modal learning scheme. The 3D deformable transformer consists of three attention modules: 3D deformability, local joint stride, and temporal stride attention. The two cross-modal tokens are input into the 3D deformable attention module to create a cross-attention token with a reflected spatiotemporal correlation. Local joint stride attention is applied to spatially combine attention and pose tokens. Temporal stride attention temporally reduces the number of input tokens in the attention module and supports temporal expression learning without the simultaneous use of all tokens. The deformable transformer iterates L-times and combines the last cross-modal token for classification. The proposed 3D deformable transformer was tested on the NTU60, NTU120, FineGYM, and PennAction datasets, and showed results better than or similar to pre-trained state-of-the-art methods even without a pre-training process. In addition, by visualizing important joints and correlations during action recognition through spatial joint and temporal stride attention, the possibility of achieving an explainable potential for action recognition is presented.
****
What do neural networks learn in image classification? A frequency shortcut perspective
Shunxin Wang, Raymond Veldhuis, Christoph Brune, Nicola Strisciuglio; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1433-1442
Frequency analysis is useful for understanding the mechanisms of representation learning in neural networks (NNs). Most research in this area focuses on the learning dynamics of NNs for regression tasks, while little for classification. This study empirically investigates the latter and expands the understanding of frequency shortcuts. First, we perform experiments on synthetic datasets, designed to have a bias in different frequency bands. Our results demonstrate that NNs tend to find simple solutions for classification, and what they learn first during training depends on the most distinctive frequency characteristics, which can be either low- or high-frequencies. Second, we confirm this phenomenon on natural images. We propose a metric to measure class-wise frequency characteristics and a method to identify frequency shortcuts. The results show that frequency shortcuts can be texture-based or shape-based, depending on what best simplifies the

objective. Third, we validate the transferability of frequency shortcuts on out-of-distribution (OOD) test sets. Our results suggest that frequency shortcuts can be transferred across datasets and cannot be fully avoided by larger model capacity and data augmentation. We recommend that future research should focus on effective training schemes mitigating frequency shortcut learning. Codes and data are available at https://github.com/nis-research/nn-frequency-shortcuts.
****
Tracking by 3D Model Estimation of Unknown Objects in Videos
Denys Rozumnyi, Ji■í Matas, Marc Pollefeys, Vittorio Ferrari, Martin R. Oswald; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14086-14096
Most model-free visual object tracking methods formulate the tracking task as object location estimation given by a 2D segmentation or a bounding box in each video frame. We argue that this representation is limited and instead propose to guide and improve 2D tracking with an explicit object representation, namely the textured 3D shape and 6DoF pose in each video frame. Our representation tackles a complex long-term dense correspondence problem between all 3D points on the object for all video frames, including frames where some points are invisible. To achieve that, the estimation is driven by re-rendering the input video frames as well as possible through differentiable rendering, which has not been used for tracking before. The proposed optimization minimizes a novel loss function to estimate the best 3D shape, texture, and 6DoF pose. We improve the state-of-the-art in 2D segmentation tracking on three different datasets with mostly rigid objects.
****
ScatterNeRF: Seeing Through Fog with Physically-Based Inverse Neural Rendering
Andrea Ramazzina, Mario Bijelic, Stefanie Walz, Alessandro Sanvito, Dominik Scheuble, Felix Heide; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17957-17968
Vision in adverse weather conditions, whether it be snow, rain, or fog is challenging. In these scenarios, scattering and attenuation severly degrades image quality. Handling such inclement weather conditions, however, is essential to operate autonomous vehicles, drones and robotic applications where human performance is impeded the most. A large body of work explores removing weather-induced image degradations with dehazing methods. Most methods rely on single images as input and struggle to generalize from synthetic fully-supervised training approaches or to generate high fidelity results from unpaired real-world datasets. With data as bottleneck and most of today's training data relying on good weather conditions with inclement weather as outlier, we rely on an inverse rendering approach to reconstruct the scene content. We introduce ScatterNeRF, a neural rendering method which adequately renders foggy scenes and decomposes the fog-free background from the participating media -- exploiting the multiple views from a short automotive sequence without the need for a large training data corpus. Instead, the rendering approach is optimized on the multi-view scene itself, which can be typically captured by an autonomous vehicle, robot or drone during operation. Specifically, we propose a disentangled representation for the scattering volume and the scene objects, and learn the scene reconstruction with physics-inspired losses. We validate our method by capturing multi-view In-the-Wild data and controlled captures in a large-scale fog chamber. Our code and datasets are available at https://light.princeton.edu/scatternerf.
****
Sigmoid Loss for Language Image Pre-Training
Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, Lucas Beyer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11975-11986
We propose a simple pairwise sigmoid loss for image-text pre-training. Unlike standard contrastive learning with softmax normalization, the sigmoid loss operates solely on image-text pairs and does not require a global view of the pairwise similarities for normalization. The sigmoid loss simultaneously allows further scaling up the batch size, while also performing better at smaller batch sizes. W

ith only four TPUv4 chips, we can train a Base CLIP model at 4k batch size and a
 Large LiT model at 20k batch size, the latter achieves 84.5% ImageNet zero-shot
 accuracy in two days. This disentanglement of the batch size from the loss furt
her allows us to study the impact of examples vs pairs and negative to positive
ratio. Finally, we push the batch size to the extreme, up to one million, and fi
nd that the benefits of growing batch size quickly diminish, with a more reasona
ble batch size of 32k being sufficient. We hope our research motivates further e
xplorations in improving the quality and efficiency of language-image pre-traini
ng.
****
PromptCap: Prompt-Guided Image Captioning for VQA with GPT-3
Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A. Smith, Jiebo Luo; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 2963-2975
Knowledge-based visual question answering (VQA) involves questions that require
world knowledge beyond the image to yield the correct answer. Large language mod
els (LMs) like GPT-3 are particularly helpful for this task because of their str
ong knowledge retrieval and reasoning capabilities. To enable LM to understand i
mages, prior work uses a captioning model to convert images into text. However,
when summarizing an image in a single caption sentence, which visual entities to
 describe are often underspecified. Generic image captions often miss visual det
ails essential for the LM to answer visual questions correctly. To address this
challenge, we propose PromptCap (Prompt-guided image Captioning), a captioning m
odel designed to serve as a better connector between images and black-box LMs. D
ifferent from generic captions, PromptCap takes a natural-language prompt to con
trol the visual entities to describe in the generated caption. The prompt contai
ns a question that the caption should aid in answering. To avoid extra annotatio
n, PromptCap is trained by examples synthesized with GPT-3 and existing datasets
. We demonstrate PromptCap's effectiveness on an existing pipeline in which GPT-
3 is prompted with image captions to carry out VQA. PromptCap outperforms generi
c captions by a large margin and achieves state-of-the-art accuracy on knowledge
-based VQA tasks (60.4% on OK-VQA and 59.6% on A-OKVQA). Zero-shot results on We
bQA show that PromptCap generalizes well to unseen domains.
****
Neural Video Depth Stabilizer
Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian,
 Guosheng Lin; Proceedings of the IEEE/CVF International Conference on Computer
Vision (ICCV), 2023, pp. 9466-9476
Video depth estimation aims to infer temporally consistent depth. Some methods a
chieve temporal consistency by finetuning a single-image depth model during test
 time using geometry and re-projection constraints, which is inefficient and not
 robust. An alternative approach is to learn how to enforce temporal consistency
 from data, but this requires well-designed models and sufficient video depth da
ta. To address these challenges, we propose a plug-and-play framework called Neu
ral Video Depth Stabilizer (NVDS) that stabilizes inconsistent depth estimations
 and can be applied to different single-image depth models without extra effort.
 We also introduce a large-scale dataset, Video Depth in the Wild (VDW), which c
onsists of 14,203 videos with over two million frames, making it the largest nat
ural-scene video depth dataset to our knowledge. We evaluate our method on the V
DW dataset as well as two public benchmarks and demonstrate significant improvem
ents in consistency, accuracy, and efficiency compared to previous approaches. O
ur work serves as a solid baseline and provides a data foundation for learning-b
ased video depth models. We will release our dataset and code for future researc
h.
****
Learning Symmetry-Aware Geometry Correspondences for 6D Object Pose Estimation
Heng Zhao, Shenxing Wei, Dahu Shi, Wenming Tan, Zheyang Li, Ye Ren, Xing Wei, Yi
 Yang, Shiliang Pu; Proceedings of the IEEE/CVF International Conference on Comp
uter Vision (ICCV), 2023, pp. 14045-14054
Current 6D pose estimation methods focus on handling objects that are previously

trained, which limits their applications in real dynamic world. To this end, we propose a geometry correspondence-based framework, termed GCPose, to estimate 6D pose of arbitrary unseen objects without any re-training. Specifically, the proposed method draws the idea from point cloud registration and resorts to object-agnostic geometry features to establish the 3D-3D correspondences between the object-scene point cloud and object-model point cloud. Then the 6D pose parameters are solved by a least-squares fitting algorithm. Taking the symmetry properties of objects into consideration, we design a symmetry-aware matching loss to facilitate the learning of dense point-wise geometry features and improve the performance considerably. Moreover, we introduce an online training data generation with special data augmentation and normalization to empower the network to learn diverse geometry prior. With training on synthetic objects from ShapeNet, our method outperforms previous approaches for unseen object pose estimation by a large margin on T-LESS, LINEMOD, Occluded-LINEMOD, and TUD-L datasets. Code is available at https://github.com/hikvision-research/GCPose.
****
TrackFlow: Multi-Object tracking with Normalizing Flows
Gianluca Mancusi, Aniello Panariello, Angelo Porrello, Matteo Fabbri, Simone Calderara, Rita Cucchiara; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9531-9543
The field of multi-object tracking has recently seen a renewed interest in the good old schema of tracking-by-detection, as its simplicity and strong priors spare it from the complex design and painful babysitting of tracking-by-attention approaches. In view of this, we aim at extending tracking-by-detection to multi-modal settings, where a comprehensive cost has to be computed from heterogeneous information e.g., 2D motion cues, visual appearance, and pose estimates. More precisely, we follow a case study where a rough estimate of 3D information is also available and must be merged with other traditional metrics (e.g., the IoU). To achieve that, recent approaches resort to either simple rules or complex heuristics to balance the contribution of each cost. However, i) they require careful tuning of tailored hyperparameters on a hold-out set, and ii) they imply these costs to be independent, which does not hold in reality. We address these issues by building upon an elegant probabilistic formulation, which considers the cost of a candidate association as the negative log-likelihood yielded by a deep density estimator, trained to model the conditional joint probability distribution of correct associations. Our experiments, conducted on both simulated and real benchmarks, show that our approach consistently enhances the performance of several tracking-by-detection algorithms.
****
Towards Generic Image Manipulation Detection with Weakly-Supervised Self-Consistency Learning
Yuanhao Zhai, Tianyu Luan, David Doermann, Junsong Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22390-22400
As advanced image manipulation techniques emerge, detecting the manipulation becomes increasingly important. Despite the success of recent learning-based approaches for image manipulation detection, they typically require expensive pixel-level annotations to train, while exhibiting degraded performance when testing on images that are differently manipulated compared with training images. To address these limitations, we propose weakly-supervised image manipulation detection, such that only binary image-level labels (authentic or tampered with) are required for training purpose. Such weakly-supervised setting can leverage more training images and has the potential to adapt quickly to new manipulation techniques. To improve the generalization ability, we propose weakly-supervised self-consistency learning (WSCL) to leverage the weakly annotated images. For the second problem, we propose an end-to-end learnable method, which takes advantage of image self-consistency properties. Specifically, two consistency properties are learned: multi-source consistency (MSC) and inter-patch consistency (IPC). MSC exploits different content-agnostic information and enables cross-source learning via an online pseudo label generation and refinement process. IPC performs global pair-wise patch-patch relationship reasoning to discover a complete region of mani

pulation. Extensive experiments validate that our WSCL, even though is weakly supervised, exhibits competitive performance compared with fully-supervised counterpart under both in-distribution and out-of-distribution evaluations, as well as reasonable manipulation localization ability.
****
PARF: Primitive-Aware Radiance Fusion for Indoor Scene Novel View Synthesis
Haiyang Ying, Baowei Jiang, Jinzhi Zhang, Di Xu, Tao Yu, Qionghai Dai, Lu Fang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17706-17716
This paper proposes a method for fast scene radiance field reconstruction with strong novel view synthesis performance and convenient scene editing functionality. The key idea is to fully utilize semantic parsing and primitive extraction for constraining and accelerating the radiance field reconstruction process. To fulfill this goal, a primitive-aware hybrid rendering strategy was proposed to enjoy the best of both volumetric and primitive rendering. We further contribute a reconstruction pipeline conducts primitive parsing and radiance field learning iteratively for each input frame which successfully fuses semantic, primitive, and radiance information into a single framework. Extensive evaluations demonstrate the fast reconstruction ability, high rendering quality, and convenient editing functionality of our method.
****
DeePoint: Visual Pointing Recognition and Direction Estimation
Shu Nakamura, Yasutomo Kawanishi, Shohei Nobuhara, Ko Nishino; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20577-20587
In this paper, we realize automatic visual recognition and direction estimation of pointing. We introduce the first neural pointing understanding method based on two key contributions. The first is the introduction of a first-of-its-kind large-scale dataset for pointing recognition and direction estimation, which we refer to as the DP Dataset. DP Dataset consists of more than 2 million frames of 33 people pointing in various styles annotated for each frame with pointing timings and 3D directions. The second is DeePoint, a novel deep network model for joint recognition and 3D direction estimation of pointing. DeePoint is a Transformer-based network which fully leverages the spatio-temporal coordination of the body parts, not just the hands. Through extensive experiments, we demonstrate the accuracy and efficiency of DeePoint. We believe DP Dataset and DeePoint will serve as a sound foundation for visual human intention understanding.
****
Periodically Exchange Teacher-Student for Source-Free Object Detection
Qipeng Liu, Luojun Lin, Zhifeng Shen, Zhifeng Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6414-6424
Source-free object detection (SFOD) aims to adapt the source detector to unlabeled target domain data in the absence of source domain data. Most SFOD methods follow the same self-training paradigm using mean-teacher (MT) framework where the student model is guided by only one single teacher model. However, such paradigm can easily fall into a training instability problem that when the teacher model collapses uncontrollably due to the domain shift, the student model also suffers drastic performance degradation. To address this issue, we propose the Periodically Exchange Teacher-Student (PETS) method, a simple yet novel approach that introduces a multiple-teacher framework consisting of a static teacher, a dynamic teacher, and a student model. During the training phase, we periodically exchange the weights between the static teacher and the student model. Then, we update the dynamic teacher using the moving average of the student model that has already been exchanged by the static teacher. In this way, the dynamic teacher can integrate knowledge from past periods, effectively reducing error accumulation and enabling a more stable training process within the MT-based framework. Further, we develop a consensus mechanism to merge the predictions of two teacher models to provide higher-quality pseudo labels for student model. Extensive experiments on multiple SFOD benchmarks show that the proposed method achieves state-of-the-art performance compared with other related methods, demonstrating the effec

tiveness and superiority of our method on SFOD task.
****
Generating Instance-level Prompts for Rehearsal-free Continual Learning
Dahuin Jung, Dongyoon Han, Jihwan Bang, Hwanjun Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11847-11857
We introduce Domain-Adaptive Prompt (DAP), a novel method for continual learning using Vision Transformers (ViT). Prompt-based continual learning has recently gained attention due to its rehearsal-free nature. Currently, the prompt pool, which is suggested by prompt-based continual learning, is key to effectively exploiting the frozen pre-trained ViT backbone in a sequence of tasks. However, we observe that the use of a prompt pool creates a domain scalability problem between pre-training and continual learning. This problem arises due to the inherent encoding of group-level instructions within the prompt pool. To address this problem, we propose DAP, a pool-free approach that generates a suitable prompt in an instance-level manner at inference time. We optimize an adaptive prompt generator that creates instance-specific fine-grained instructions required for each input, enabling enhanced model plasticity and reduced forgetting. Our experiments on seven datasets with varying degrees of domain similarity to ImageNet demonstrate the superiority of DAP over state-of-the-art prompt-based methods. Code is publicly available at https://github.com/naver-ai/dap-cl.
****
Deformer: Dynamic Fusion Transformer for Robust Hand Pose Estimation
Qichen Fu, Xingyu Liu, Ran Xu, Juan Carlos Niebles, Kris M. Kitani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23600-23611
Accurately estimating 3D hand pose is crucial for understanding how humans interact with the world. Despite remarkable progress, existing methods often struggle to generate plausible hand poses when the hand is heavily occluded or blurred. In videos, the movements of the hand allow us to observe various parts of the hand that may be occluded or blurred in a single frame. To adaptively leverage the visual clue before and after the occlusion or blurring for robust hand pose estimation, we propose the Deformer: a framework that implicitly reasons about the relationship between hand parts within the same image (spatial dimension) and different timesteps (temporal dimension). We show that a naive application of the transformer self-attention mechanism is not sufficient because motion blur or occlusions in certain frames can lead to heavily distorted hand features and generate imprecise keys and queries. To address this challenge, we incorporate a Dynamic Fusion Module into Deformer, which predicts the deformation of the hand and warps the hand mesh predictions from nearby frames to explicitly support the current frame estimation. Furthermore, we have observed that errors are unevenly distributed across different hand parts, with vertices around fingertips having disproportionately higher errors than those around the palm. We mitigate this issue by introducing a new loss function called maxMSE that automatically adjusts the weight of every vertex to focus the model on critical hand parts. Extensive experiments show that our method significantly outperforms state-of-the-art methods by 10%, and is more robust to occlusions (over 14%).
****
HSE: Hybrid Species Embedding for Deep Metric Learning
Bailin Yang, Haoqiang Sun, Frederick W. B. Li, Zheng Chen, Jianlu Cai, Chao Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11047-11057
Deep metric learning is crucial for finding an embedding function that can generalize to training and testing data, including unknown test classes. However, limited training samples restrict the model's generalization to downstream tasks. While adding new training samples is a promising solution, determining their labels remains a significant challenge. Here, we introduce Hybrid Species Embedding (HSE), which employs mixed sample data augmentations to generate hybrid species and provide additional training signals. We demonstrate that HSE outperforms multiple state-of-the-art methods in improving the metric Recall@K on the CUB-200 , CAR-196 and SOP datasets, thus offering a novel solution to deep metric learnin

g's limitations.
****
Online Continual Learning on Hierarchical Label Expansion
Byung Hyun Lee, Okchul Jung, Jonghyun Choi, Se Young Chun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11761-11770
Continual learning (CL) enables models to adapt to new tasks and environments without forgetting previously learned knowledge. While current CL setups have ignored the relationship between labels in the past task and the new task with or without small task overlaps, real-world scenarios often involve hierarchical relationships between old and new tasks, posing another challenge for traditional CL approaches. To address this challenge, we propose a novel multi-level hierarchical class incremental task configuration with an online learning constraint, called hierarchical label expansion (HLE). Our configuration allows a network to first learn coarse-grained classes, with data labels continually expanding to more fine-grained classes in various hierarchy depths. To tackle this new setup, we propose a rehearsal-based method that utilizes hierarchy-aware pseudo-labeling to incorporate hierarchical class information. Additionally, we propose a simple yet effective memory management and sampling strategy that selectively adopts samples of newly encountered classes. Our experiments demonstrate that our proposed method can effectively use hierarchy on our HLE setup to improve classification accuracy across all levels of hierarchies, regardless of depth and class imbalance ratio, outperforming prior state-of-the-art works by significant margins while also outperforming them on the conventional disjoint, blurry and i-Blurry CL setups.
****
iDAG: Invariant DAG Searching for Domain Generalization
Zenan Huang, Haobo Wang, Junbo Zhao, Nenggan Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19169-19179
Existing machine learning (ML) models are often fragile in open environments because the data distribution frequently shifts. To address this problem, domain generalization (DG) aims to explore underlying invariant patterns for stable prediction across domains. In this work, we first characterize that this failure of conventional ML models in DG is attributed to an inadequate identification of causal structures. We further propose a novel and theoretically grounded invariant Directed Acyclic Graph (dubbed iDAG) searching framework that attains an invariant graphical relation as the proxy to the causality structure from the intrinsic data-generating process. To enable tractable computation, iDAG solves a constrained optimization objective built on a set of representative class-conditional prototypes. Additionally, we integrate a hierarchical contrastive learning module, which poses a strong effect of clustering, for enhanced prototypes as well as stabler prediction. Extensive experiments on the synthetic and real-world benchmarks demonstrate that iDAG outperforms the state-of-the-art approaches, verifying the superiority of causal structure identification for DG. The code of iDAG is available at https://github.com/lccurious/iDAG.
****
Spacetime Surface Regularization for Neural Dynamic Scene Reconstruction
Jaesung Choe, Christopher Choy, Jaesik Park, In So Kweon, Anima Anandkumar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17871-17881
We propose an algorithm, 4DRegSDF, for the spacetime surface regularization to improve the fidelity of neural rendering and reconstruction in dynamic scenes. The key idea is to impose local rigidity on the deformable Signed Distance Function (SDF) for temporal coherency. Our approach works by (1) sampling points on the deformed surface by taking gradient steps toward the steepest direction along SDF, (2) extracting differential surface geometry, such as tangent plane or curvature, at each sample, and (3) adjusting the local rigidity at different timestamps. This enables our dynamic surface regularization to align 4D spacetime geometry via 3D canonical space more accurately. Experiments demonstrate that our 4DRegSDF achieves state-of-the-art performance in both reconstruction and rendering quality over synthetic and real-world datasets.

****
GasMono: Geometry-Aided Self-Supervised Monocular Depth Estimation for Indoor Scenes
Chaoqiang Zhao, Matteo Poggi, Fabio Tosi, Lei Zhou, Qiyu Sun, Yang Tang, Stefano Mattoccia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16209-16220
This paper tackles the challenges of self-supervised monocular depth estimation in indoor scenes caused by large rotation between frames and low texture. We ease the learning process by obtaining coarse camera poses from monocular sequences through multi-view geometry to deal with the former. However, we found that limited by the scale ambiguity across different scenes in the training dataset, a naive introduction of geometric coarse poses cannot play a positive role in performance improvement, which is counter-intuitive.
To address this problem, we propose to refine those poses during training through rotation and translation/scale optimization.
To soften the effect of the low texture, we combine the global reasoning of vision transformers with an overfitting-aware, iterative self-distillation mechanism, providing more accurate depth guidance coming from the network itself.
Experiments on NYUv2, ScanNet, 7scenes, and KITTI datasets support the effectiveness of each component in our framework, which sets a new state-of-the-art for indoor self-supervised monocular depth estimation, as well as outstanding generalization ability. Code and models are available at https://github.com/zxcqlf/GasMono
****
3D Motion Magnification: Visualizing Subtle Motions from Time-Varying Radiance Fields
Brandon Y. Feng, Hadi Alzayer, Michael Rubinstein, William T. Freeman, Jia-bin Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9837-9846
Motion magnification helps us visualize subtle, imperceptible motion. However, prior methods only work for 2D videos captured with a fixed camera. We present a 3D motion magnification method that can magnify subtle motions from scenes captured by a moving camera, while supporting novel view rendering. We represent the scene with time-varying radiance fields and leverage the Eulerian principle for motion magnification to extract and amplify the variation of the embedding of a fixed point over time. We study and validate our proposed principle for 3D motion magnification using both implicit and tri-plane-based radiance fields as our underlying 3D scene representation. We evaluate the effectiveness of our method on both synthetic and real-world scenes captured under various camera setups.
****
Learning to Transform for Generalizable Instance-wise Invariance
Utkarsh Singhal, Carlos Esteves, Ameesh Makadia, Stella X. Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6211-6221
Computer vision research has long aimed to build systems that are robust to transformations found in natural data. Traditionally, this is done using data augmentation or hard-coding invariances into the architecture.
However, too much or too little invariance can hurt, and the correct amount is unknown a priori and dependent on the instance. Ideally, the appropriate invariance would be learned from data and inferred at test-time.

We treat invariance as a prediction problem. Given any image, we predict a distribution over transformations. We use variational inference to learn this distribution end-to-end. Combined with a graphical model approach, this distribution forms a flexible, generalizable, and adaptive form of invariance. Our experiments show that it can be used to align datasets and discover prototypes, adapt to out-of-distribution poses, and generalize invariances across classes. When used for data augmentation, our method shows consistent gains in accuracy and robustness on CIFAR 10, CIFAR10-LT, and TinyImageNet.
****

Audio-Visual Deception Detection: DOLOS Dataset and Parameter-Efficient Crossmodal Learning

Xiaobao Guo, Nithish Muthuchamy Selvaraj, Zitong Yu, Adams Wai-Kin Kong, Bingquan Shen, Alex Kot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22135-22145

Deception detection in conversations is a challenging yet important task, having pivotal applications in many fields such as credibility assessment in business, multimedia anti-frauds, and custom security. Despite this, deception detection research is hindered by the lack of high-quality deception datasets, as well as the difficulties of learning multimodal features effectively. To address this issue, we introduce DOLOS, the largest gameshow deception detection dataset with rich deceptive conversations. DOLOS includes 1,675 video clips featuring 213 subjects, and it has been labeled with audio-visual feature annotations. We provide train-test, duration, and gender protocols to investigate the impact of different factors. We benchmark our dataset on previously proposed deception detection approaches. To further improve the performance by fine-tuning fewer parameters, we propose Parameter-Efficient Crossmodal Learning (PECL), where a Uniform Temporal Adapter (UT-Adapter) explores temporal attention in transformer-based architectures, and a crossmodal fusion module, Plug-in Audio-Visual Fusion (PAVF), combines crossmodal information from audio-visual features. Based on the rich fine-grained audio-visual annotations on DOLOS, we also exploit multi-task learning to enhance performance by concurrently predicting deception and audio-visual features. Experimental results demonstrate the desired quality of the DOLOS dataset and the effectiveness of the PECL. The DOLOS dataset and the source codes are available.

****

Multiple Instance Learning Framework with Masked Hard Instance Mining for Whole Slide Image Classification

Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, Bo Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4078-4087

The whole slide image (WSI) classification is often formulated as a multiple instance learning (MIL) problem. Since the positive tissue is only a small fraction of the gigapixel WSI, existing MIL methods intuitively focus on identifying salient instances via attention mechanisms. However, this leads to a bias towards easy-to-classify instances while neglecting hard-to-classify instances. Some literature has revealed that hard examples are beneficial for modeling a discriminative boundary accurately. By applying such an idea at the instance level, we elaborate a novel MIL framework with masked hard instance mining (MHIM-MIL), which uses a Siamese structure (Teacher-Student) with a consistency constraint to explore the potential hard instances. With several instance masking strategies based on attention scores, MHIM-MIL employs a momentum teacher to implicitly mine hard instances for training the student model, which can be any attention-based MIL model. This counter-intuitive strategy essentially enables the student to learn a better discriminating boundary. Moreover, the student is used to update the teacher with an exponential moving average (EMA), which in turn identifies new hard instances for subsequent training iterations and stabilizes the optimization. Experimental results on the CAMELYON-16 and TCGA Lung Cancer datasets demonstrate that MHIM-MIL outperforms other latest methods in terms of performance and training cost. The code is available at: https://github.com/DearCaat/MHIM-MIL.

****

Unsupervised Compositional Concepts Discovery with Text-to-Image Generative Models

Nan Liu, Yilun Du, Shuang Li, Joshua B. Tenenbaum, Antonio Torralba; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2085-2095

Text-to-image generative models have enabled high-resolution image synthesis across different domains, but require users to specify the content they wish to generate. In this paper, we consider the inverse problem - given a collection of different images, can we discover the generative concepts that represent each imag

e? We present an unsupervised approach to discover generative concepts from a co
llection of images, disentangling different art styles in paintings, objects, an
d lighting from kitchen scenes, and discovering image classes given ImageNet ima
ges. We show how such generative concepts can accurately represent the content o
f images, be recombined and composed to generate new artistic and hybrid images,
 and be further used as a representation for downstream classification tasks.
****
Partition-And-Debias: Agnostic Biases Mitigation via a Mixture of Biases-Specifi
c Experts
Jiaxuan Li, Duc Minh Vo, Hideki Nakayama; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2023, pp. 4924-4934
Bias mitigation in image classification has been widely researched, and existing
 methods have yielded notable results. However, most of these methods implicitly
 assume that a given image contains only one type of known or unknown bias, fail
ing to consider the complexities of real-world biases. We introduce a more chall
enging scenario, agnostic biases mitigation, aiming at bias removal regardless o
f whether the type of bias or the number of types is unknown in the datasets. To
 address this difficult task, we present the Partition-and-Debias (PnD) method t
hat uses a mixture of biases-specific experts to implicitly divide the bias spac
e into multiple subspaces and a gating module to find a consensus among experts
to achieve debiased classification. Experiments on both public and constructed b
enchmarks demonstrated the efficacy of the PnD. Code is available at: https://gi
thub.com/Jiaxuan-Li/PnD.
****
Spatial Self-Distillation for Object Detection with Inaccurate Bounding Boxes
Di Wu, Pengfei Chen, Xuehui Yu, Guorong Li, Zhenjun Han, Jianbin Jiao; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
 6855-6865
Object detection via inaccurate bounding box supervision has boosted a broad int
erest due to the expensive high-quality annotation data or the occasional inevit
ability of low annotation quality (e.g. tiny objects). The previous works usuall
y utilize multiple instance learning (MIL), which highly depends on category inf
ormation, to select and refine a low-quality box. Those methods suffer from part
 domination, object drift and group prediction problems without exploring spatia
l information. In this paper, we heuristically propose a Spatial Self-Distillati
on based Object Detector (SSD-Det) to mine spatial information to refine the ina
ccurate box in a self-distillation fashion. SSD-Det utilizes a Spatial Position
Self-Distillation SPSD) module to exploit spatial information and an interactive
 structure to combine spatial information and category information, thus constru
cting a high-quality proposal bag. To further improve the selection procedure, a
 Spatial Identity Self-Distillation (SISD) module is introduced in SSD-Det to ob
tain spatial confidence to help select the best proposals. Experiments on MS-COC
O and VOC datasets with noisy box annotation verify our method's effectiveness a
nd achieve state-of-the-art performance. The code is available at https://github
.com/ucas-vg/PointTinyBenchmark/tree/SSD-Det.
****
CC3D: Layout-Conditioned Generation of Compositional 3D Scenes
Sherwin Bahmani, Jeong Joon Park, Despoina Paschalidou, Xingguang Yan, Gordon We
tzstein, Leonidas Guibas, Andrea Tagliasacchi; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2023, pp. 7171-7181
In this work, we introduce CC3D, a conditional generative model that synthesizes
 complex 3D scenes conditioned on 2D semantic scene layouts, trained using singl
e-view images. Different from most existing 3D GANs that limit their applicabili
ty to aligned single objects, we focus on generating complex scenes with multipl
e objects, by modeling the compositional nature of 3D scenes. By devising a 2D l
ayout-based approach for 3D synthesis and implementing a new 3D field representa
tion with a stronger geometric inductive bias, we have created a 3D GAN that is
both efficient and of high quality, while allowing for a more controllable gener
ation process. Our evaluations on synthetic 3D-FRONT and real-world KITTI-360 da
tasets demonstrate that our model generates scenes of improved visual and geomet

ric quality in comparison to previous works.
****
Alleviating Catastrophic Forgetting of Incremental Object Detection via Within-Class and Between-Class Knowledge Distillation
Mengxue Kang, Jinpeng Zhang, Jinming Zhang, Xiashuang Wang, Yang Chen, Zhe Ma, Xuhui Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18894-18904
Incremental object detection (IOD) task requires a model to learn continually from newly added data. However, directly fine-tuning a well-trained detection model on a new task will sharply decrease the performance on old tasks, which is known as catastrophic forgetting. Knowledge distillation, including feature distillation and response distillation, has been proven to be an effective way to alleviate catastrophic forgetting. However, previous works on feature distillation heavily rely on low-level feature information, while under-exploring the importance of high-level semantic information. In this paper, we discuss the cause of catastrophic forgetting in IOD task as destruction of semantic feature space. We propose a method that dynamically distills both semantic and feature information with consideration of both between-class discriminativeness and within-class consistency on Transformer-based detector. Between-class discriminativeness is preserved by distilling class-level semantic distance and feature distance among various categories, while within-class consistency is preserved by distilling instance-level semantic information and feature information within each category. Extensive experiments are conducted on both Pascal VOC and MS COCO benchmarks. Our method outperforms all the previous CNN-based SOTA methods under various experimental scenarios, with a remarkable mAP improvement from 36.90% to 39.80% under one-step IOD task.
****
TextPSG: Panoptic Scene Graph Generation from Textual Descriptions
Chengyang Zhao, Yikang Shen, Zhenfang Chen, Mingyu Ding, Chuang Gan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2839-2850
Panoptic Scene Graph has recently been proposed for comprehensive scene understanding. However, previous works adopt a fully-supervised learning manner, requiring large amounts of pixel-wise densely-annotated data, which is always tedious and expensive to obtain. To address this limitation, we study a new problem of Panoptic Scene Graph Generation from Purely Textual Descriptions (Caption-to-PSG). The key idea is to leverage the large collection of free image-caption data on the Web alone to generate panoptic scene graphs. The problem is very challenging for three constraints: 1) no location priors; 2) no explicit links between visual regions and textual entities; and 3) no pre-defined concept sets. To tackle this problem, we propose a new framework TextPSG consisting of four modules, i.e., a region grouper, an entity grounder, a segment merger, and a label generator, with several novel techniques. The region grouper first groups image pixels into different segments and the entity grounder then aligns visual segments with language entities based on the textual description of the segment being referred to. The grounding results can thus serve as pseudo labels enabling the segment merger to learn the segment similarity as well as guiding the label generator to learn object semantics and relation predicates, resulting in a fine-grained structured scene understanding. Our framework is effective, significantly outperforming the baselines and achieving strong out-of-distribution robustness. We perform comprehensive ablation studies to corroborate the effectiveness of our design choices and provide an in-depth analysis to highlight future directions. Our code, data, and results are available on our project page: https://vis-www.cs.umass.edu/TextPSG.
****
Revisiting the Parameter Efficiency of Adapters from the Perspective of Precision Redundancy
Shibo Jie, Haoqing Wang, Zhi-Hong Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17217-17226
Current state-of-the-art results in computer vision depend in part on fine-tunin

g large pre-trained vision models. However, with the exponential growth of model sizes, the conventional full fine-tuning, which needs to store a individual network copy for each tasks, leads to increasingly huge storage and transmission overhead. Adapter-based Parameter-Efficient Tuning (PET) methods address this challenge by tuning lightweight adapters inserted into the frozen pre-trained models. In this paper, we investigate how to make adapters even more efficient, reaching a new minimum size required to store a task-specific fine-tuned network. Inspired by the observation that the parameters of adapters converge at flat local minima, we find that adapters are resistant to noise in parameter space, which means they are also resistant to low numerical precision. To train low-precision adapters, we propose a computational-efficient quantization method which minimizes the quantization error. Through extensive experiments, we find that low-precision adapters exhibit minimal performance degradation, and even 1-bit precision is sufficient for adapters. The results of the experiments demonstrate that 1-bit adapters outperform all other PET methods on both the VTAB-1K benchmark and few-shot FGVC datasets, while requiring the smallest storage size. Our findings show, for the first time, the significant potential of quantization techniques in PET, providing a general solution to enhance the parameter efficiency of adapter-based PET methods.

****

EMQ: Evolving Training-free Proxies for Automated Mixed Precision Quantization

Peijie Dong, Lujun Li, Zimian Wei, Xin Niu, Zhiliang Tian, Hengyue Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17076-17086

Mixed-Precision Quantization (MQ) can achieve a competitive accuracy-complexity trade-off for models. Conventional training-based search methods require time-consuming candidate training to search optimized per-layer bit-width configurations in MQ. Recently, some training-free approaches have presented various MQ proxies and significantly improve search efficiency. However, the correlation between these proxies and quantization accuracy is poorly understood. To address the gap, we first build the MQ-Bench-101, which involves different bit configurations and quantization results. Then, we observe that the existing training-free proxies perform weak correlations on the MQ-Bench-101. To efficiently seek superior proxies, we develop an automatic search of proxies framework for MQ via evolving algorithms. In particular, we devise an elaborate search space involving the existing proxies and perform an evolution search to discover the best correlated MQ proxy. We proposed a diversity-prompting selection strategy and compatibility screening protocol to avoid premature convergence and improve search efficiency. In this way, our Evolving proxies for Mixed-precision Quantization (EMQ) framework allows the auto-generation of proxies without heavy tuning and expert knowledge. Extensive experiments on ImageNet with various ResNet and MobileNet families demonstrate that our EMQ obtains superior performance than state-of-the-art mixed-precision methods at a significantly reduced cost. The code will be released.

****

Face Clustering via Graph Convolutional Networks with Confidence Edges

Yang Wu, Zhiwei Ge, Yuhao Luo, Lin Liu, Sulong Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20990-20999

Face clustering is a method for unlabeled image annotation and has attracted increasing attention. Existing methods have made significant breakthroughs by introducing Graph Convolutional Networks (GCNs) on the affinity graph. However, such graphs will contain many vertex pairs with inconsistent similarities and labels, thus degrading the model's performance. There are already relevant efforts for this problem, but the information about features needs to be mined further. In this paper, we define a new concept called confidence edge and guide the construction of graphs. Furthermore, a novel confidence-GCN is proposed to cluster face images by deriving more confidence edges. Firstly, Local Information Fusion is advanced to obtain a more accurate similarity metric by considering the neighbors of vertices. Then Unsupervised Neighbor Determination is used to discard low-quality edges based on similarity differences. Moreover, we elaborate that the remaining edges retain the most beneficial information to demonstrate the validity.

At last, the confidence-GCN takes the graph as the input and fully uses the con fidence edges to complete the clustering. Experiments show that our method outpe rforms existing methods on the face and person datasets to achieve state-of-the-art. At the same time, comparable results are obtained on the fashion dataset.
****
Learning Spatial-context-aware Global Visual Feature Representation for Instance Image Retrieval
Zhongyan Zhang, Lei Wang, Luping Zhou, Piotr Koniusz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11250-11259
In instance image retrieval, considering local spatial information within an ima ge has proven effective to boost retrieval performance, as demonstrated by local visual descriptor based geometric verification. Nevertheless, it will be highly valuable to make ordinary global image representations spatial-context-aware be cause global representation based image retrieval is appealing thanks to its alg orithmic simplicity, low memory cost, and being friendly to sophisticated data s tructures. To this end, we propose a novel feature learning framework for instan ce image retrieval, which embeds local spatial context information into the lear ned global feature representations. Specifically, in parallel to the visual feat ure branch in a CNN backbone, we design a spatial context branch that consists o f two modules called online token learning and distance encoding. For each local descriptor learned in CNN, the former module is used to indicate the types of i ts surrounding descriptors, while their spatial distribution information is capt ured by the latter module. After that, the visual feature branch and the spatial context branch are fused to produce a single global feature representation per image. As experimentally demonstrated, with the spatial-context-aware characteri stic, we can well improve the performance of global representation based image r etrieval while maintaining all of its appealing properties. Our code is availabl e at https://github.com/Zy-Zhang/SpCa
****
Cross-modal Latent Space Alignment for Image to Avatar Translation
Manuel Ladron de Guevara, Jose Echevarria, Yijun Li, Yannick Hold-Geoffroy, Came ron Smith, Daichi Ito; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2023, pp. 520-529
We present a novel method for automatic vectorized avatar generation from a sing le portrait image. Most existing approaches that create avatars rely on image-to -image translation methods, which present some limitations when applied to 3D re ndering, animation, or video. Instead, we leverage modality-specific autoencoder s trained on large-scale unpaired portraits and parametric avatars, and then lea rn a mapping between both modalities via an alignment module trained on a signif icantly smaller amount of data. The resulting cross-modal latent space preserves facial identity, producing more visually appealing and higher fidelity avatars than previous methods, as supported by our quantitative and qualitative evaluati ons. Moreover, our method's virtue of being resolution-independent makes it high ly versatile and applicable in a wide range of settings.
****
Inspecting the Geographical Representativeness of Images from Text-to-Image Mode ls
Abhipsa Basu, R. Venkatesh Babu, Danish Pruthi; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2023, pp. 5136-5147
Recent progress in generative models has resulted in models that produce both re alistic as well as relevant images for most textual inputs. These models are bei ng used to generate millions of images everyday, and hold the potential to drast ically impact areas such as generative art, digital marketing and data augmentat ion. Given their outsized impact, it is important to ensure that the generated c ontent reflects the artifacts and surroundings across the globe, rather than ove r-representing certain parts of the world. In this paper, we measure the geograp hical representativeness of common nouns (e.g., a house) generated through DALL.E 2 and Stable Diffusion models using a crowdsourced study comprising 540 partic ipants across 27 countries. For deliberately underspecified inputs without count ry names, the generated images most reflect the surroundings of the United State

s followed by India, and the top generations rarely reflect surroundings from all other countries (average score less than 3 out of 5). Specifying the country names in the input increases the representativeness by 1.44 points on average on a 5-point Likert scale for DALL.E 2 and 0.75 for Stable Diffusion, however, the overall scores for many countries still remain low, highlighting the need for future models to be more geographically inclusive. Lastly, we examine the feasibility of quantifying the geographical representativeness of generated images without conducting user studies.
****
Space-time Prompting for Video Class-incremental Learning
Yixuan Pei, Zhiwu Qing, Shiwei Zhang, Xiang Wang, Yingya Zhang, Deli Zhao, Xueming Qian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11932-11942
Recently, prompt-based learning has made impressive progress on image class-incremental learning, but it still lacks sufficient exploration in the video domain. In this paper, we will fill this gap by learning multiple prompts based on a powerful image-language pre-trained model, i.e., CLIP, making it fit for video class-incremental learning (VCIL). For this purpose, we present a space-time prompting approach (ST-Prompt) which contains two kinds of prompts, i.e., task-specific prompts and task-agnostic prompts. The task-specific prompts are to address the catastrophic forgetting problem by learning multi-grained prompts, i.e., spatial prompts, temporal prompts and comprehensive prompts, for accurate task identification. The task-agnostic prompts maintain a globally-shared prompt pool, which can empower the pre-trained image models with temporal perception abilities by exchanging contexts between frames. By this means, ST-Prompt can transfer the plentiful knowledge in the image-language pre-trained models to the VCIL task with only a tiny set of prompts to be optimized. To evaluate ST-Prompt, we conduct extensive experiments on three standard benchmarks. The results show that ST-Prompt can significantly surpass the state-of-the-art VCIL methods, especially it gains 9.06% on HMDB51 dataset under the 1*25 stage setting.
****
Multimodal Garment Designer: Human-Centric Latent Diffusion Models for Fashion Image Editing
Alberto Baldrati, Davide Morelli, Giuseppe Cartella, Marcella Cornia, Marco Bertini, Rita Cucchiara; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23393-23402
Fashion illustration is used by designers to communicate their vision and to bring the design idea from conceptualization to realization, showing how clothes interact with the human body. In this context, computer vision can thus be used to improve the fashion design process. Differently from previous works that mainly focused on the virtual try-on of garments, we propose the task of multimodal-conditioned fashion image editing, guiding the generation of human-centric fashion images by following multimodal prompts, such as text, human body poses, and garment sketches. We tackle this problem by proposing a new architecture based on latent diffusion models, an approach that has not been used before in the fashion domain. Given the lack of existing datasets suitable for the task, we also extend two existing fashion datasets, namely Dress Code and VITON-HD, with multimodal annotations collected in a semi-automatic manner. Experimental results on these new datasets demonstrate the effectiveness of our proposal, both in terms of realism and coherence with the given multimodal inputs. Source code and collected multimodal annotations are publicly available at: https://github.com/aimagelab/multimodal-garment-designer.
****
Time-to-Contact Map by Joint Estimation of Up-to-Scale Inverse Depth and Global Motion using a Single Event Camera
Urbano Miguel Nunes, Laurent Udo Perrinet, Sio-Hoi Ieng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23653-23663
Event cameras asynchronously report brightness changes with a temporal resolution in the order of microseconds, which makes them inherently suitable to address problems that involve rapid motion perception. In this paper, we address the pro

blem of time-to-contact (TTC) estimation using a single event camera. This probl
em is typically addressed by estimating a single global TTC measure, which expli
citly assumes that the surface/obstacle is planar and fronto-parallel. We relax
this assumption by proposing an incremental event-based method to estimate the T
TC that jointly estimates the (up-to scale) inverse depth and global motion usin
g a single event camera. The proposed method is reliable and fast while asynchro
nously maintaining a TTC map (TTCM), which provides per-pixel TTC estimates. As
a side product, the proposed method can also estimate per-event optical flow. We
 achieve state-of-the-art performances on TTC estimation in terms of accuracy an
d runtime per event while achieving competitive performance on optical flow esti
mation.
****
Sparse Sampling Transformer with Uncertainty-Driven Ranking for Unified Removal
of Raindrops and Rain Streaks
Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, Lei Zhu; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1310
6-13117
In the real world, image degradations caused by rain often exhibit a combination
 of rain streaks and raindrops, thereby increasing the challenges of recovering
the underlying clean image. Note that the rain streaks and raindrops have divers
e shapes, sizes, and locations in the captured image, and thus modeling the corr
elation relationship between irregular degradations caused by rain artifacts is
a necessary prerequisite for image deraining. This paper aims to present an effi
cient and flexible mechanism to learn and model degradation relationships in a g
lobal view, thereby achieving a unified removal of intricate rain scenes. To do
so, we propose a Sparse Sampling Transformer based on Uncertainty-Driven Ranking
, dubbed UDR-S2Former. Compared to previous methods, our UDR-S2Former has three
merits. First, it can adaptively sample relevant image degradation information t
o model underlying degradation relationships. Second, explicit application of th
e uncertainty-driven ranking strategy can facilitate the network to attend to de
gradation features and understand the reconstruction process. Finally, experimen
tal results show that our UDR-S2Former clearly outperforms state-of-the-art meth
ods for all benchmarks.
****
A Benchmark for Chinese-English Scene Text Image Super-Resolution
Jianqi Ma, Zhetong Liang, Wangmeng Xiang, Xi Yang, Lei Zhang; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19452-19
461
Scene Text Image Super-resolution (STISR) aims to recover high-resolution (HR) s
cene text images with visually pleasant and readable text content from the given
 low-resolution (LR) input. Most existing works focus on recovering English text
s, which have simple structures in the characters, while little work has been do
ne on the more challenging Chinese texts with diverse and complex character stru
ctures. In this paper, we propose a real-world Chinese-English benchmark dataset
, namely Real-CE, for the task of STISR with the emphasis on restoring structura
lly complex Chinese characters. The benchmark provides 1,935/783 real-world LR-H
R text image pairs (contains 33,789 text lines in total) for training/testing in
 2x and 4x zooming modes, complemented by detailed annotations, including detect
ion boxes and text transcripts. Moreover, we design an edge-aware learning metho
d, which provides structural supervision in image and feature domain, to effecti
vely reconstruct the dense structures of Chinese characters. We conduct experime
nts on the proposed Real-CE benchmark and evaluate the existing STISR models wit
h and without our edge-aware loss. The benchmark, including data and source code
, will be made publicly available.
****
HSR-Diff: Hyperspectral Image Super-Resolution via Conditional Diffusion Models
Chanyue Wu, Dong Wang, Yunpeng Bai, Hanyu Mao, Ying Li, Qiang Shen; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 70
83-7093
Despite the proven significance of hyperspectral images (HSIs) in performing var

ious computer vision tasks, its potential is adversely affected by the low-resol ution (LR) property in the spatial domain, resulting from multiple physical fact ors. Inspired by recent advancements in deep generative models, we propose an HS I Super-resolution (SR) approach with Conditional Diffusion Models (HSR-Diff) th at merges a high-resolution (HR) multispectral image (MSI) with the correspondin g LR-HSI. HSR-Diff generates an HR-HSI via repeated refinement, in which the HR-HSI is initialized with pure Gaussian noise and iteratively refined. At each ite ration, the noise is removed with a Conditional Denoising Transformer (CDFormer) that is trained on denoising at different noise levels, conditioned on the hier archical feature maps of HR-MSI and LR-HSI. In addition, a progressive learning strategy is employed to exploit the global information of full-resolution images . Systematic experiments have been conducted on four public datasets, demonstrat ing that HSR-Diff outperforms state-of-the-art methods.
****
Replay: Multi-modal Multi-view Acted Videos for Casual Holography
Roman Shapovalov, Yanir Kleiman, Ignacio Rocco, David Novotny, Andrea Vedaldi, C hangan Chen, Filippos Kokkinos, Ben Graham, Natalia Neverova; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20338-20 348
We introduce Replay, a collection of multi-view, multi-modal videos of humans in teracting socially. Each scene is filmed in high production quality, from differ ent viewpoints with several static cameras, as well as wearable action cameras, and recorded with a large array of microphones at different positions in the roo m. Overall, the dataset contains over 3000 minutes of footage and over 5 million timestamped high-resolution frames annotated with camera poses and partially wi th foreground masks. The Replay dataset has many potential applications, such as novel-view synthesis, 3D reconstruction, novel-view acoustic synthesis, human b ody and face analysis, and training generative models. We provide a benchmark fo r training and evaluating novel-view synthesis, with two scenarios of different difficulty. Finally, we evaluate several baseline state-of-the-art methods on th e new benchmark.
****
Advancing Example Exploitation Can Alleviate Critical Challenges in Adversarial Training
Yao Ge, Yun Li, Keji Han, Junyi Zhu, Xianzhong Long; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 145-154
Deep neural networks have achieved remarkable results across various tasks. Howe ver, they are susceptible to adversarial examples, which are generated by adding adversarial perturbations to original data. Adversarial training (AT) is the mo st effective defense mechanism against adversarial examples and has received sig nificant attention. Recent studies highlight the importance of example exploitat ion, where the model's learning intensity is altered for specific examples to ex tend classic AT approaches. However, the analysis methodologies employed by thes e studies are varied and contradictory, which may lead to confusion in future re search. To address this issue, we provide a comprehensive summary of representat ive strategies focusing on exploiting examples within a unified framework. Furth ermore, we investigate the role of examples in AT and find that examples which c ontribute primarily to accuracy or robustness are distinct. Based on this findin g, we propose a novel example-exploitation idea that can further improve the per formance of advanced AT methods. This new idea suggests that critical challenges in AT, such as the accuracy-robustness trade-off, robust overfitting, and catas trophic overfitting, can be alleviated simultaneously from an example-exploitati on perspective. The code can be found in https://github.com/geyao1995/advancing-example-exploitation-in-adversarial-training.
****
Affine-Consistent Transformer for Multi-Class Cell Nuclei Detection
Junjia Huang, Haofeng Li, Xiang Wan, Guanbin Li; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2023, pp. 21384-21393
Multi-class cell nuclei detection is a fundamental prerequisite in the diagnosis of histopathology. It is critical to efficiently locate and identify cells with

diverse morphology and distributions in digital pathological images. Most existing methods take complex intermediate representations as learning targets and rely on inflexible post-refinements while paying less attention to various cell density and fields of view. In this paper, we propose a novel Affine-Consistent Transformer (AC-Former), which directly yields a sequence of nucleus positions and is trained collaboratively through two sub-networks, a global and a local network. The local branch learns to infer distorted input images of smaller scales while the global network outputs the large-scale predictions as extra supervision signals. We further introduce an Adaptive Affine Transformer (AAT) module, which can automatically learn the key spatial transformations to warp original images for local network training. The AAT module works by learning to capture the transformed image regions that are more valuable for training the model. Experimental results demonstrate that the proposed method significantly outperforms existing state-of-the-art algorithms on various benchmarks.
****

Removing Anomalies as Noises for Industrial Defect Localization
Fanbin Lu, Xufeng Yao, Chi-Wing Fu, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16166-16175
Unsupervised anomaly detection aims to train models with only anomaly-free images to detect and localize unseen anomalies. Previous reconstruction-based methods have been limited by inaccurate reconstruction results. This work presents a denoising model to detect and localize the anomalies with a generative diffusion model. In particular, we introduce random noise to overwhelm the anomalous pixels and obtain pixel-wise precise anomaly scores from the intermediate denoising process. We find that the KL divergence of the diffusion model serves as a better anomaly score compared with the traditional RGB space score. Furthermore, we reconstruct the features from a pre-trained deep feature extractor as our feature level score to improve localization performance. Moreover, we propose a gradient denoising process to smoothly transform an anomalous image into a normal one. Our denoising model outperforms the state-of-the-art reconstruction-based anomaly detection methods for precise anomaly localization and high-quality normal image reconstruction on the MVTec-AD benchmark.
****

GPGait: Generalized Pose-based Gait Recognition
Yang Fu, Shibei Meng, Saihui Hou, Xuecai Hu, Yongzhen Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19595-19604
Recent works on pose-based gait recognition have demonstrated the potential of using such simple information to achieve results comparable to silhouette-based methods. However, the generalization ability of pose-based methods on different datasets is undesirably inferior to that of silhouette-based ones, which has received little attention but hinders the application of these methods in real-world scenarios. To improve the generalization ability of pose-based methods across datasets, we propose a Generalized Pose-based Gait recognition (GPGait) framework. First, a Human-Oriented Transformation (HOT) and a series of Human-Oriented Descriptors (HOD) are proposed to obtain a unified pose representation with discriminative multi-features. Then, given the slight variations in the unified representation after HOT and HOD, it becomes crucial for the network to extract local-global relationships between the keypoints. To this end, a Part-Aware Graph Convolutional Network (PAGCN) is proposed to enable efficient graph partition and local-global spatial feature extraction. Experiments on four public gait recognition datasets, CASIA-B, OUMVLP-Pose, Gait3D and GREW, show that our model demonstrates better and more stable cross-domain capabilities compared to existing skeleton-based methods, achieving comparable recognition results to silhouette-based ones. Code is available at https://github.com/BNU-IVC/FastPoseGait.
****

Stable and Causal Inference for Discriminative Self-supervised Deep Visual Representations
Yuewei Yang, Hai Li, Yiran Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16109-16120

In recent years, discriminative self-supervised methods have made significant strides in advancing various visual tasks. The central idea of learning a data encoder that is robust to data distortions/augmentations is straightforward yet highly effective. Although many studies have demonstrated the empirical success of various learning methods, the resulting learned representations can exhibit instability and hinder downstream performance. In this study, we analyze discriminative self-supervised methods from a causal perspective to explain these unstable behaviors and propose solutions to overcome them. Our approach draws inspiration from prior works that empirically demonstrate the ability of discriminative self-supervised methods to demix ground truth causal sources to some extent. Unlike previous work on causality-empowered representation learning, we do not apply our solutions during the training process but rather during the inference process to improve time efficiency. Through experiments on both controlled image datasets and realistic image datasets, we show that our proposed solutions, which involve tempering a linear transformation with controlled synthetic data, are effective in addressing these issues.
****
ShiftNAS: Improving One-shot NAS via Probability Shift
Mingyang Zhang, Xinyi Yu, Haodong Zhao, Linlin Ou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5919-5928
One-shot Neural architecture search (One-shot NAS) has been proposed as a time-efficient approach to obtain optimal subnet architectures and weights under different complexity cases by training only once. However, the subnet performance obtained by weight sharing is often inferior to the performance achieved by retraining. In this paper, we investigate the performance gap and attribute it to the use of uniform sampling, which is a common approach in supernet training. Uniform sampling concentrates training resources on subnets with intermediate computational resources, which are sampled with high probability. However, subnets with different complexity regions require different optimal training strategies for optimal performance.
To address the problem of uniform sampling, we propose ShiftNAS, a method that can adjust the sampling probability based on the complexity of subnets. We achieve this by evaluating the performance variation of subnets with different complexity and designing an architecture generator that can accurately and efficiently provide subnets with the desired complexity. Both the sampling probability and the architecture generator can be trained end-to-end in a gradient-based manner. With ShiftNAS, we can directly obtain the optimal model architecture and parameters for a given computational complexity. We evaluate our approach on multiple visual network models, including convolutional neural networks (CNNs) and vision transformers (ViTs), and demonstrate that ShiftNAS is model-agnostic. Experimental results on ImageNet show that ShiftNAS can improve the performance of one-shot NAS without additional computational consumption. Source codes are available at GitHub.
****
Semantic Attention Flow Fields for Monocular Dynamic Scene Decomposition
Yiqing Liang, Eliot Laidlaw, Alexander Meyerowitz, Srinath Sridhar, James Tompkin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21797-21806
From video, we reconstruct a neural volume that captures time-varying color, density, scene flow, semantics, and attention information. The semantics and attention let us identify salient foreground objects separately from the background across spacetime. To mitigate low resolution semantic and attention features, we compute pyramids that trade detail with whole-image context. After optimization, we perform a saliency-aware clustering to decompose the scene. To evaluate real-world scenes, we annotate object masks in the NVIDIA Dynamic Scene and DyCheck datasets. We demonstrate that this method can decompose dynamic scenes in an unsupervised way with competitive performance to a supervised method, and that it improves foreground/background segmentation over recent static/dynamic split methods. Project webpage: https://visual.cs.brown.edu/saff
****

LexLIP: Lexicon-Bottlenecked Language-Image Pre-Training for Large-Scale Image-Text Sparse Retrieval
Ziyang Luo, Pu Zhao, Can Xu, Xiubo Geng, Tao Shen, Chongyang Tao, Jing Ma, Qingwei Lin, Daxin Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11206-11217
Image-text retrieval (ITR) aims to retrieve images or texts that match a query originating from the other modality. The conventional dense retrieval paradigm relies on encoding images and texts into dense representations with dual-stream encoders. However, this approach is limited by slow retrieval speeds in large-scale scenarios. To address this issue, we propose a novel sparse retrieval paradigm for ITR that exploits sparse representations in the vocabulary space for images and texts. This paradigm enables us to leverage bag-of-words models and efficient inverted indexes, significantly reducing retrieval latency. A critical gap emerges from representing continuous image data in a sparse vocabulary space. To bridge this gap, we introduce a novel pre-training framework, Lexicon-Bottlenecked Language-Image Pre-Training (LexLIP), that learns importance-aware lexicon representations. By using lexicon-bottlenecked modules between the dual-stream encoders and weakened text decoders, we are able to construct continuous bag-of-words bottlenecks and learn lexicon-importance distributions. Upon pre-training with same-scale data, our LexLIP achieves state-of-the-art performance on two ITR benchmarks, MSCOCO and Flickr30k. Furthermore, in large-scale retrieval scenarios, LexLIP outperforms CLIP with 5.8x faster retrieval speed and 19.1x less index storage memory. Beyond this, LexLIP surpasses CLIP across 8 out of 10 zero-shot image classification tasks.
****
A Fast Unified System for 3D Object Detection and Tracking
Thomas Heitzinger, Martin Kampel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17044-17054
We present FUS3D, a fast and lightweight system for real-time 3D object detection and tracking on edge devices. Our approach seamlessly integrates stages for 3D object detection and multi-object-tracking into a single, end-to-end trainable model. FUS3D is specially tuned for indoor 3D human behavior analysis, with target applications in Ambient Assisted Living (AAL) or surveillance. The system is optimized for inference on the edge, thus enabling sensor-near processing of potentially sensitive data. In addition, our system relies exclusively on the less privacy-intrusive 3D depth imaging modality, thus further highlighting the potential of our method for application in sensitive areas. FUS3D achieves best results when utilized in a joint detection and tracking configuration. Nevertheless, the proposed detection stage can function as a fast standalone object detection model if required. We have evaluated FUS3D extensively on the MIPT dataset and demonstrated its superior performance over comparable existing state-of-the-art methods in terms of 3D object detection, multi-object tracking, and most importantly, runtime. Model code will be made publicly available.
****
Adaptive Testing of Computer Vision Models
Irena Gao, Gabriel Ilharco, Scott Lundberg, Marco Tulio Ribeiro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4003-4014
Vision models often fail systematically on groups of data that share common semantic characteristics (e.g., rare objects or unusual scenes), but identifying these failure modes is a challenge. We introduce AdaVision, an interactive process for testing vision models which helps users identify and fix coherent failure modes. Given a natural language description of a coherent group, AdaVision retrieves relevant images from LAION-5B with CLIP. The user then labels a small amount of data for model correctness, which is used in successive retrieval rounds to hill-climb towards high-error regions, refining the group definition. Once a group is saturated, AdaVision uses GPT-3 to suggest new group descriptions for the user to explore. We demonstrate the usefulness and generality of AdaVision in user studies, where users find major bugs in state-of-the-art classification, object detection, and image captioning models. These user-discovered groups have fail

ure rates 2-3x higher than those surfaced by automatic error clustering methods. Finally, finetuning on examples found with AdaVision fixes the discovered bugs when evaluated on unseen examples, without degrading in-distribution accuracy, and while also improving performance on out-of-distribution datasets.
****
LFS-GAN: Lifelong Few-Shot Image Generation
Juwon Seo, Ji-Su Kang, Gyeong-Moon Park; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11356-11366
We address a challenging lifelong few-shot image generation task for the first time. In this situation, a generative model learns a sequence of tasks using only a few samples per task. Consequently, the learned model encounters both catastrophic forgetting and overfitting problems at a time. Existing studies on lifelong GANs have proposed modulation-based methods to prevent catastrophic forgetting. However, they require considerable additional parameters and cannot generate high-fidelity and diverse images from limited data. On the other hand, the existing few-shot GANs suffer from severe catastrophic forgetting when learning multiple tasks. To alleviate these issues, we propose a framework called Lifelong Few-Shot GAN (LFS-GAN) that can generate high-quality and diverse images in lifelong few-shot image generation task. Our proposed framework learns each task using an efficient task-specific modulator - Learnable Factorized Tensor (LeFT). LeFT is rank-constrained and has a rich representation ability due to its unique reconstruction technique. Furthermore, we propose a novel mode seeking loss to improve the diversity of our model in low-data circumstances. Extensive experiments demonstrate that the proposed LFS-GAN can generate high-fidelity and diverse images without any forgetting and mode collapse in various domains, achieving state-of-the-art in lifelong few-shot image generation task. Surprisingly, we find that our LFS-GAN even outperforms the existing few-shot GANs in the few-shot image generation task. The code is available at Github.
****
AIDE: A Vision-Driven Multi-View, Multi-Modal, Multi-Tasking Dataset for Assistive Driving Perception
Dingkang Yang, Shuai Huang, Zhi Xu, Zhenpeng Li, Shunli Wang, Mingcheng Li, Yuzheng Wang, Yang Liu, Kun Yang, Zhaoyu Chen, Yan Wang, Jing Liu, Peixuan Zhang, Peng Zhai, Lihua Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20459-20470
Driver distraction has become a significant cause of severe traffic accidents over the past decade. Despite the growing development of vision-driven driver monitoring systems, the lack of comprehensive perception datasets restricts road safety and traffic security. In this paper, we present an AssIstive Driving pErception dataset (AIDE) that considers context information both inside and outside the vehicle in naturalistic scenarios. AIDE facilitates holistic driver monitoring through three distinctive characteristics, including multi-view settings of driver and scene, multi-modal annotations of face, body, posture, and gesture, and four pragmatic task designs for driving understanding. To thoroughly explore AIDE, we provide experimental benchmarks on three kinds of baseline frameworks via extensive methods. Moreover, two fusion strategies are introduced to give new insights into learning effective multi-stream/modal representations. We also systematically investigate the importance and rationality of the key components in AIDE and benchmarks. The project link is https://github.com/ydk122024/AIDE.
****
Feature Proliferation -- the "Cancer" in StyleGAN and its Treatments
Shuang Song, Yuanbang Liang, Jing Wu, Yu-Kun Lai, Yipeng Qin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2360-2370
Despite the success of StyleGAN in image synthesis, the images it synthesizes are not always perfect and the well-known truncation trick has become a standard post-processing technique for StyleGAN to synthesize high-quality images. Although effective, it has long been noted that the truncation trick tends to reduce the diversity of synthesized images and unnecessarily sacrifices many distinct image features. To address this issue, in this paper, we first delve into the Style

GAN image synthesis mechanism and discover an important phenomenon, namely Feature Proliferation, which demonstrates how specific features reproduce with forward propagation. Then, we show how the occurrence of Feature Proliferation results in StyleGAN image artifacts. As an analogy, we refer to it as the "cancer" in StyleGAN from its proliferating and malignant nature. Finally, we propose a novel feature rescaling method that identifies and modulates risky features to mitigate feature proliferation. Thanks to our discovery of Feature Proliferation, the proposed feature rescaling method is less destructive and retains more useful image features than the truncation trick, as it is more fine-grained and works in a lower-level feature space rather than a high-level latent space. Experimental results justify the validity of our claims and the effectiveness of the proposed feature rescaling method. Our code is available at https://github.com/songc42/Feature-proliferation.
****
Self-Supervised Character-to-Character Distillation for Text Recognition
Tongkun Guan, Wei Shen, Xue Yang, Qi Feng, Zekun Jiang, Xiaokang Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19473-19484
When handling complicated text images (e.g., irregular structures, low resolution, heavy occlusion, and uneven illumination), existing supervised text recognition methods are data-hungry. Although these methods employ large-scale synthetic text images to reduce the dependence on annotated real images, the domain gap still limits the recognition performance. Therefore, exploring the robust text feature representations on unlabeled real images by self-supervised learning is a good solution. However, existing self-supervised text recognition methods conduct sequence-to-sequence representation learning by roughly splitting the visual features along the horizontal axis, which limits the flexibility of the augmentations, as large geometric-based augmentations may lead to sequence-to-sequence feature inconsistency. Motivated by this, we propose a novel self-supervised Character-to-Character Distillation method, CCD, which enables versatile augmentations to facilitate general text representation learning. Specifically, we delineate the character structures of unlabeled real images by designing a self-supervised character segmentation module. Following this, CCD easily enriches the diversity of local characters while keeping their pairwise alignment under flexible augmentations, using the transformation matrix between two augmented views from images. Experiments demonstrate that CCD achieves state-of-the-art results, with average performance gains of 1.38% in text recognition, 1.7% in text segmentation, 0.24 dB (PSNR) and 0.0321 (SSIM) in text super-resolution. Code will be released soon.
****
MixCycle: Mixup Assisted Semi-Supervised 3D Single Object Tracking with Cycle Consistency
Qiao Wu, Jiaqi Yang, Kun Sun, Chu'ai Zhang, Yanning Zhang, Mathieu Salzmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13956-13966
3D single object tracking (SOT) is an indispensable part of automated driving. Existing approaches rely heavily on large, densely labeled datasets. However, annotating point clouds is both costly and time-consuming. Inspired by the great success of cycle tracking in unsupervised 2D SOT, we introduce the first semi-supervised approach to 3D SOT. Specifically, we introduce two cycle-consistency strategies for supervision: 1) Self tracking cycles, which leverage labels to help the model converge better in the early stages of training; 2) forward-backward cycles, which strengthen the tracker's robustness to motion variations and the template noise caused by the template update strategy. Furthermore, we propose a data augmentation strategy named SOTMixup to improve the tracker's robustness to point cloud diversity. SOTMixup generates training samples by sampling points in two point clouds with a mixing rate and assigns a reasonable loss weight for training according to the mixing rate. The resulting MixCycle approach generalizes to appearance matching-based trackers. On the KITTI benchmark, based on the P2B tracker, MixCycle trained with 10% labels outperforms P2B trained with 100% labe

ls, and achieves a 28.4% precision improvement when using 1% labels. Our code wi
ll be released at https://github.com/Mumuqiao/MixCycle.
****
Multi-Label Self-Supervised Learning with Scene Images
Ke Zhu, Minghao Fu, Jianxin Wu; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 6694-6703
Self-supervised learning (SSL) methods targeting scene images have seen a rapid
growth recently, and they mostly rely on either a dedicated dense matching mecha
nism or a costly unsupervised object discovery module. This paper shows that ins
tead of hinging on these strenuous operations, quality image representations can
 be learned by treating scene/multi-label image SSL simply as a multi-label clas
sification problem, which greatly simplifies the learning framework. Specificall
y, multiple binary pseudo-labels are assigned for each input image by comparing
its embeddings with those in two dictionaries, and the network is optimized usin
g the binary cross entropy loss. The proposed method is named Multi-Label Self-s
upervised learning (MLS). Visualizations qualitatively show that clearly the pse
udo-labels by MLS can automatically find semantically similar pseudo-positive pa
irs across different images to facilitate contrastive learning. MLS learns high
quality representations on MS-COCO and achieves state-of-the-art results on clas
sification, detection and segmentation benchmarks. At the same time, MLS is much
 simpler than existing methods, making it easier to deploy and for further explo
ration.
****
Domain Adaptive Few-Shot Open-Set Learning
Debabrata Pal, Deeptej More, Sai Bhargav, Dipesh Tamboli, Vaneet Aggarwal, Bipla
b Banerjee; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2023, pp. 18831-18840
Few-shot learning has made impressive strides in addressing the crucial challeng
es of recognizing unknown samples from novel classes in target query sets and ma
naging visual shifts between domains. However, existing techniques fall short wh
en it comes to identifying target outliers under domain shifts by learning to re
ject pseudo-outliers from the source domain, resulting in an incomplete solution
 to both problems. To address these challenges comprehensively, we propose a nov
el approach called Domain Adaptive Few-Shot Open Set Recognition (DA-FSOS) and i
ntroduce a meta-learning-based architecture named DAFOS-Net. During training, ou
r model learns a shared and discriminative embedding space while creating a pseu
do-open-space decision boundary, given a fully-supervised source domain and a la
bel-disjoint few-shot target domain. To enhance data density, we use a pair of c
onditional adversarial networks with tunable noise variances to augment both dom
ains' closed and pseudo-open spaces. Furthermore, we propose a domain-specific b
atch-normalized class prototypes alignment strategy to align both domains global
ly while ensuring class-discriminativeness through novel metric objectives. Our
training approach ensures that DAFOS-Net can generalize well to new scenarios in
 the target domain. We present three benchmarks for DA-FSOS based on the Office-
Home, mini-ImageNet/CUB, and DomainNet datasets and demonstrate the efficacy of
DAFOS-Net through extensive experimentation.
****
DiffFacto: Controllable Part-Based 3D Point Cloud Generation with Cross Diffusio
n
George Kiyohiro Nakayama, Mikaela Angelina Uy, Jiahui Huang, Shi-Min Hu, Ke Li,
Leonidas Guibas; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 14257-14267
While the community of 3D point cloud generation has witnessed a big growth in r
ecent years, there still lacks an effective way to enable intuitive user control
 in the generation process, hence limiting the general utility of such methods.
Since an intuitive way of decomposing a shape is through its parts, we propose t
o tackle the task of controllable part-based point cloud generation. We introduc
e DiffFacto, a novel probabilistic generative model that learns the distribution
 of shapes with part-level control. We propose a factorization that models indep
endent part style and part configuration distributions, and present a novel cros

s diffusion network that enables us to generate coherent and plausible shapes under our proposed factorization. Experiments show that our method is able to generate novel shapes with multiple axes of control. It achieves state-of-the-art part-level generation quality and generates plausible and coherent shape, while enabling various downstream editing applications such as shape interpolation, mixing and transformation editing. Code will be made publicly available.
****
Interactive Class-Agnostic Object Counting
Yifeng Huang, Viresh Ranjan, Minh Hoai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22312-22322
We propose a novel framework for interactive class-agnostic object counting, where a human user can interactively provide feedback to improve the accuracy of a counter. Our framework consists of two main components: a user-friendly visualizer to gather feedback and an efficient mechanism to incorporate it. In each iteration, we produce a density map to show the current prediction result, and we segment it into non-overlapping regions with an easily verifiable number of objects. The user can provide feedback by selecting a region with obvious counting errors and specifying the range for the estimated number of objects within it. To improve the counting result, we develop a novel adaptation loss to force the visual counter to output the predicted count within the user-specified range. For effective and efficient adaptation, we propose a refinement module that can be used with any density-based visual counter, and only the parameters in the refinement module will be updated during adaptation. Our experiments on two challenging class-agnostic object counting benchmarks, FSCD-LVIS and FSC-147, show that our method can reduce the mean absolute error of multiple state-of-the-art visual counters by roughly 30% to 40% with minimal user input. Our project can be found at https://yifehuang97.github.io/ICACountProjectPage/.
****
Spatio-temporal Prompting Network for Robust Video Feature Extraction
Guanxiong Sun, Chi Wang, Zhaoyu Zhang, Jiankang Deng, Stefanos Zafeiriou, Yang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13587-13597
The frame quality deterioration problem is one of the main challenges in the field of video understanding. To compensate for the information loss caused by deteriorated frames, recent approaches exploit transformer-based integration modules to obtain spatio-temporal information. However, these integration modules are heavy and complex. Furthermore, each integration module is specifically tailored for its target task, making it difficult to generalise to multiple tasks. In this paper, we present a neat and unified framework, called Spatio-Temporal Prompting Network (STPN). It can efficiently extract robust and accurate video features by dynamically adjusting the input features in the backbone network. Specifically, STPN predicts several video prompts containing spatio-temporal information of neighbour frames. Then, these video prompts are prepended to the patch embeddings of the current frame as the updated input for video feature extraction. Moreover, STPN is easy to generalise to various video tasks because it does not contain task-specific modules. Without bells and whistles, STPN achieves state-of-the-art performance on three widely-used datasets for different video understanding tasks, i.e., ImageNetVID for video object detection, YouTubeVIS for video instance segmentation, and GOT-10k for visual object tracking. Codes are available at https://github.com/guanxiongsun/STPN.
****
Enhancing Fine-Tuning Based Backdoor Defense with Sharpness-Aware Minimization
Mingli Zhu, Shaokui Wei, Li Shen, Yanbo Fan, Baoyuan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4466-4477
Backdoor defense, which aims to detect or mitigate the effect of malicious triggers introduced by attackers, is becoming increasingly critical for machine learning security and integrity. Fine-tuning based on benign data is a natural defense to erase the backdoor effect in a backdoored model. However, recent studies show that, given limited benign data, vanilla fine-tuning has poor defense performance. In this work, we firstly investigate the vanilla fine-tuning process for b

ackdoor mitigation from the neuron weight perspective, and find that backdoor-related neurons are only slightly perturbed in the vanilla fine-tuning process, which explains its poor backdoor defense performance. To enhance the fine-tuning based defense, inspired by the observation that the backdoor-related neurons often have larger weight norms, we propose FT-SAM, a novel backdoor defense paradigm that aims to shrink the norms of backdoorrelated neurons by incorporating sharpness-aware minimization with fine-tuning. We demonstrate the effectiveness of our method on several benchmark datasets and network architectures, where it achieves state-of-the-art defense performance, and provide extensive analysis to reveal the FTSAM's mechanism. Overall, our work provides a promising avenue for improving the robustness of machine learning models against backdoor attacks. Codes are available at https://github.com/SCLBD/BackdoorBench.
****
Deep Geometry-Aware Camera Self-Calibration from Video
Annika Hagemann, Moritz Knorr, Christoph Stiller; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3438-3448
Accurate intrinsic calibration is essential for camera-based 3D perception, yet, it typically requires targets of well-known geometry. Here, we propose a camera self-calibration approach that infers camera intrinsics during application, from monocular videos in the wild. We propose to explicitly model projection functions and multi-view geometry, while leveraging the capabilities of deep neural networks for feature extraction and matching. To achieve this, we build upon recent research on integrating bundle adjustment into deep learning models, and introduce a self-calibrating bundle adjustment layer. The self-calibrating bundle adjustment layer optimizes camera intrinsics through classical Gauss-Newton steps and can be adapted to different camera models without re-training. As a specific realization, we implemented this layer within the deep visual SLAM system DROID-SLAM, and show that the resulting model, DroidCalib, yields state-of-the-art calibration accuracy across multiple public datasets. Our results suggest that the model generalizes to unseen environments and different camera models, including significant lens distortion. Thereby, the approach enables performing 3D perception tasks without prior knowledge about the camera. Code is available at https://github.com/boschresearch/droidcalib.
****
A Simple Vision Transformer for Weakly Semi-supervised 3D Object Detection
Dingyuan Zhang, Dingkang Liang, Zhikang Zou, Jingyu Li, Xiaoqing Ye, Zhe Liu, Xiao Tan, Xiang Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8373-8383
Advanced 3D object detection methods usually rely on large-scale, elaborately labeled datasets to achieve good performance. However, labeling the bounding boxes for the 3D objects is difficult and expensive. Although semi-supervised (SS3D) and weakly-supervised 3D object detection (WS3D) methods can effectively reduce the annotation cost, they suffer from two limitations: 1) their performance is far inferior to the fully-supervised counterparts; 2) they are difficult to adapt to different detectors or scenes (e.g, indoor or outdoor). In this paper, we study weakly semi-supervised 3D object detection (WSS3D) with point annotations, where the dataset comprises a small number of fully labeled and massive weakly labeled data with a single point annotated for each 3D object. To fully exploit the point annotations, we employ the plain and non-hierarchical vision transformer to form a point-to-box converter, termed ViT-WSS3D. By modeling global interactions between LiDAR points and corresponding weak labels, our ViT-WSS3D can generate high-quality pseudo-bounding boxes, which are then used to train any 3D detectors without exhaustive tuning. Extensive experiments on indoor and outdoor datasets (SUN RGBD and KITTI) show the effectiveness of our method. In particular, when only using 10% fully labeled and the rest as point labeled data, our ViT-WSS3D can enable most detectors to achieve similar performance with the oracle model using 100% fully labeled data.
****
Estimator Meets Equilibrium Perspective: A Rectified Straight Through Estimator for Binary Neural Networks Training

Xiao-Ming Wu, Dian Zheng, Zuhao Liu, Wei-Shi Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17055-17064
Binarization of neural networks is a dominant paradigm in neural networks compression. The pioneering work BinaryConnect uses Straight Through Estimator (STE) to mimic the gradients of the sign function, but it also causes the crucial inconsistency problem. Most of the previous methods design different estimators instead of STE to mitigate it. However, they ignore the fact that when reducing the estimating error, the gradient stability will decrease concomitantly. These highly divergent gradients will harm the model training and increase the risk of gradient vanishing and gradient exploding. To fully take the gradient stability into consideration, we present a new perspective to the BNNs training, regarding it as the equilibrium between the estimating error and the gradient stability. In this view, we firstly design two indicators to quantitatively demonstrate the equilibrium phenomenon. In addition, in order to balance the estimating error and the gradient stability well, we revise the original straight through estimator and propose a power function based estimator, Rectified Straight Through Estimator (ReSTE for short). Comparing to other estimators, ReSTE is rational and capable of flexibly balancing the estimating error with the gradient stability. Extensive experiments on CIFAR-10 and ImageNet datasets show that ReSTE has excellent performance and surpasses the state-of-the-art methods without any auxiliary modules or losses.
****
Exploring Object-Centric Temporal Modeling for Efficient Multi-View 3D Object Detection
Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, Xiangyu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3621-3631
In this paper, we propose a long-sequence modeling framework, named StreamPETR, for multi-view 3D object detection. Built upon the sparse query design in the PETR series, we systematically develop an object-centric temporal mechanism. The model is performed in an online manner and the long-term historical information is propagated through object queries frame by frame. Besides, we introduce a motion-aware layer normalization to model the movement of the objects. StreamPETR achieves significant performance improvements only with negligible computation cost, compared to the single-frame baseline. On the standard nuScenes benchmark, it is the first online multi-view method that achieves comparable performance (67.6% NDS & 65.3% AMOTA) with lidar-based methods. The lightweight version realizes 45.0% mAP and 31.7 FPS, outperforming the state-of-the-art method (SOLOFusion) by 2.3% mAP and 1.8x faster FPS. Code has been available at https://github.com/exiawsh/StreamPETR.git.
****
Open-domain Visual Entity Recognition: Towards Recognizing Millions of Wikipedia Entities
Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, Ming-Wei Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12065-12075
Large-scale multi-modal pre-training models such as CLIP and PaLI exhibit strong generalization on various visual domains and tasks. However, existing image classification benchmarks often evaluate recognition on a specific domain (e.g., outdoor images) or a specific task (e.g., classifying plant species), which falls short of evaluating whether pre-trained foundational models are universal visual recognizers. To address this, we formally present the task of Open-domain Visual Entity recognitioN (OVEN), where a model need to link an image onto a Wikipedia entity with respect to a text query. We construct OVEN by re-purposing 14 existing datasets with all labels grounded onto one single label space: Wikipedia entities. OVEN challenges models to select among six million possible Wikipedia entities, making it a general visual recognition benchmark with largest number of labels. Our study on state-of-the-art pre-trained models reveals large headroom in generalizing to the massive-scale label space. We show that a PaLI-based auto-regressive visual recognition model performs surprisingly well, even on Wikiped

ia entities that have never been seen during fine-tuning. We also find existing pre-trained models yield different unique strengths: while PaLI-based models obtains higher overall performance, CLIP-based models are better at recognizing tail entities.
****
MedKLIP: Medical Knowledge Enhanced Language-Image Pre-Training for X-ray Diagnosis
Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21372-21383
In this paper, we consider enhancing medical visual-language pre-training (VLP) with domain-specific knowledge, by exploiting the paired image-text reports from the radiological daily practice. In particular, we make the following contributions: First, unlike existing works that directly process the raw reports, we adopt a novel triplet extraction module to extract the medical-related information, avoiding unnecessary complexity from language grammar and enhancing the supervision signals; Second, we propose a novel triplet encoding module with entity translation
by querying a knowledge base, to exploit the rich domain knowledge in medical field, and implicitly build relationships between medical entities in the language embedding space; Third, we propose to use a Transformer-based fusion model for spatially aligning the entity description with visual signals at the image patch level, enabling the ability for medical diagnosis; Fourth, we conduct thorough experiments to validate the effectiveness of our architecture, and benchmark on numerous public benchmarks e.g., ChestX-ray14, RSNA Pneumonia, SIIM-ACR Pneumothorax, COVIDx CXR-2, COVID Rural, and EdemaSeverity. In both zero-shot and fine-tuning settings, our model has demonstrated strong performance compared with the former methods on disease classification and grounding.
****
Automated Knowledge Distillation via Monte Carlo Tree Search
Lujun Li, Peijie Dong, Zimian Wei, Ya Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17413-17424
In this paper, we present Auto-KD, the first automated search framework for optimal knowledge distillation design. Traditional distillation techniques typically require handcrafted designs by experts and extensive tuning costs for different teacher-student pairs. To address these issues, we empirically study different distillers, finding that they can be decomposed, combined, and simplified. Based on these observations, we build our uniform search space with advanced operations in transformations, distance functions, and hyperparameters components. For instance, the transformation parts are optional for global, intra-spatial, and inter-spatial operations, such as attention, mask, and multi-scale. Then, we introduce an effective search strategy based on the Monte Carlo tree search, modeling the search space as a Monte Carlo Tree (MCT) to capture the dependency among options. The MCT is updated using test loss and representation gap of student trained by candidate distillers as the reward for better exploration-exploitation balance. To accelerate the search process, we exploit offline processing without teacher inference, sparse training for student, and proxy settings based on distillation properties. In this way, our Auto-KD only needs small costs to search for optimal distillers before the distillation phase. Moreover, we expand Auto-KD for multi-layer and multi-teacher scenarios with training-free weighted factors. Our method is promising yet practical, and extensive experiments demonstrate that it generalizes well to different CNNs and Vision Transformer models and attains state-of-the-art performance across a range of vision tasks, including image classification, object detection, and semantic segmentation. Code is provided at https://github.com/lilujunai/Auto-KD.
****
EmoTalk: Speech-Driven Emotional Disentanglement for 3D Face Animation
Ziqiao Peng, Haoyu Wu, Zhenbo Song, Hao Xu, Xiangyu Zhu, Jun He, Hongyan Liu, Zhaoxin Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20687-20697

Speech-driven 3D face animation aims to generate realistic facial expressions that match the speech content and emotion. However, existing methods often neglect emotional facial expressions or fail to disentangle them from speech content. To address this issue, this paper proposes an end-to-end neural network to disentangle different emotions in speech so as to generate rich 3D facial expressions. Specifically, we introduce the emotion disentangling encoder (EDE) to disentangle the emotion and content in the speech by cross-reconstructed speech signals with different emotion labels. Then an emotion-guided feature fusion decoder is employed to generate a 3D talking face with enhanced emotion. The decoder is driven by the disentangled identity, emotional, and content embeddings so as to generate controllable personal and emotional styles. Finally, considering the scarcity of the 3D emotional talking face data, we resort to the supervision of facial blendshapes, which enables the reconstruction of plausible 3D faces from 2D emotional data, and contribute a large-scale 3D emotional talking face dataset (3D-ETF) to train the network. Our experiments and user studies demonstrate that our approach outperforms state-of-the-art methods and exhibits more diverse facial movements. We recommend watching the supplementary video: https://ziqiaopeng. github.io/emotalk
****
A Soft Nearest-Neighbor Framework for Continual Semi-Supervised Learning
Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, Karteek Alahari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11868-11877
Despite significant advances, the performance of state-of-the-art continual learning approaches hinges on the unrealistic scenario of fully labeled data. In this paper, we tackle this challenge and propose an approach for continual semi-supervised learning--a setting where not all the data samples are labeled. A primary issue in this scenario is the model forgetting representations of unlabeled data and overfitting the labeled samples. We leverage the power of nearest-neighbor classifiers to nonlinearly partition the feature space and flexibly model the underlying data distribution thanks to its non-parametric nature. This enables the model to learn a strong representation for the current task, and distill relevant information from previous tasks. We perform a thorough experimental evaluation and show that our method outperforms all the existing approaches by large margins, setting a solid state of the art on the continual semi-supervised learning paradigm. For example, on CIFAR-100 we surpass several others even when using at least 30 times less supervision (0.8% vs. 25% of annotations). Finally, our method works well on both low and high resolution images and scales seamlessly to more complex datasets such as ImageNet-100.
****
Text-Conditioned Sampling Framework for Text-to-Image Generation with Masked Generative Models
Jaewoong Lee, Sangwon Jang, Jaehyeong Jo, Jaehong Yoon, Yunji Kim, Jin-Hwa Kim, Jung-Woo Ha, Sung Ju Hwang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23252-23262
Token-based masked generative models are gaining popularity for their fast inference time with parallel decoding. While recent token-based approaches achieve competitive performance to diffusion-based models, their generation performance is still suboptimal as they sample multiple tokens simultaneously without considering the dependence among them. We empirically investigate this problem and propose a learnable sampling model, Text-Conditioned Token Selection (TCTS), to select optimal tokens via localized supervision with text information. TCTS improves not only the image quality but also the semantic alignment of the generated images with the given texts. To further improve the image quality, we introduce a cohesive sampling strategy, Frequency Adaptive Sampling (FAS), to each group of tokens divided according to the self-attention maps. We validate the efficacy of TCTS combined with FAS with various generative tasks, demonstrating that it significantly outperforms the baselines in image-text alignment and image quality. Our text-conditioned sampling framework further reduces the original inference time by more than 50% without modifying the original generative model.

****
ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes
Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, Angela Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12-22
We present ScanNet++, a large-scale dataset that couples together capture of high-quality and commodity-level geometry and color of indoor scenes. Each scene is captured with a high-end laser scanner at sub-millimeter resolution, along with registered 33-megapixel images from a DSLR camera, and RGB-D streams from an iPhone. Scene reconstructions are further annotated with an open vocabulary of semantics, with label-ambiguous scenarios explicitly annotated for comprehensive semantic understanding. ScanNet++ enables a new real-world benchmark for novel view synthesis, both from high-quality RGB capture, and importantly also from commodity-level images, in addition to a new benchmark for 3D semantic scene understanding that comprehensively encapsulates diverse and ambiguous semantic labeling scenarios. Currently, ScanNet++ contains 460 scenes, 280,000 captured DSLR images, and over 3.7M iPhone RGBD frames.
****
Minimal Solutions to Uncalibrated Two-view Geometry with Known Epipoles
Gaku Nakano; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13361-13370
This paper proposes minimal solutions to uncalibrated two-view geometry with known epipoles. Exploiting the epipoles, we can reduce the number of point correspondences needed to find the fundamental matrix together with the intrinsic parameters: the focal length and the radial lens distortion. We define four cases by the number of available epipoles and unknown intrinsic parameters, then derive a closed-form solution for each case formulated as a higher-order polynomial in a single variable. The proposed solvers are more numerically stable and faster by orders of magnitude than the conventional 6- or 7-point algorithms. Moreover, we demonstrate by experiments on the human pose dataset that the proposed method can solve two-view geometry even with 2D human pose, of which point localization is noisier than general feature point detectors.
****
Improving Diversity in Zero-Shot GAN Adaptation with Semantic Variations
Seogkyu Jeon, Bei Liu, Pilhyeon Lee, Kibeom Hong, Jianlong Fu, Hyeran Byun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7258-7267
Training deep generative models usually requires a large amount of data. To alleviate the data collection cost, the task of zero-shot GAN adaptation aims to reuse well-trained generators to synthesize images of an unseen target domain without any further training samples. Due to the data absence, the textual description of the target domain and the vision-language models, e.g., CLIP, are utilized to effectively guide the generator. However, with only a single representative text feature instead of real images, the synthesized images gradually lose diversity as the model is optimized, which is also known as mode collapse. To tackle the problem, we propose a novel method to find semantic variations of the target text in the CLIP space. Specifically, we explore diverse semantic variations based on the informative text feature of the target domain while regularizing the uncontrolled deviation of the semantic information. With the obtained variations, we design a novel directional moment loss that matches the first and second moments of image and text direction distributions. Moreover, we introduce elastic weight consolidation and a relation consistency loss to effectively preserve valuable content information from the source domain, e.g., appearances. Through extensive experiments, we demonstrate the efficacy of the proposed methods in ensuring sample diversity in various scenarios of zero-shot GAN adaptation. We also conduct ablation studies to validate the effect of each proposed component. Notably, our model achieves a new state-of-the-art on zero-shot GAN adaptation in terms of both diversity and quality.
****
Context-Aware Planning and Environment-Aware Memory for Instruction Following Embodied Agents

Byeonghwi Kim, Jinyeon Kim, Yuyeong Kim, Cheolhong Min, Jonghyun Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10936-10946

Accomplishing household tasks such as 'bringing a cup of water' requires to plan step-by-step actions by maintaining the knowledge about the spatial arrangement of objects and consequences of previous actions. Perception models of current embodied AI agents, however, often make mistakes due to lack of such knowledge but rely on imperfect learning of imitating agents or an algorithmic planner without the knowledge about the changed environment by the previous actions. To address the issue, we propose the CPEM (Context-aware Planner and Environment-aware Memory) embodied agent to incorporate the contextual information of previous actions for planning and maintaining spatial arrangement of objects with their states (e.g., if an object has been already moved or not) in the environment to the perception model for improving both visual navigation and object interactions. We observe that the proposed model achieves state-of-the-art task success performance in various metrics using a challenging interactive instruction following benchmark both in seen and unseen environments by large margins (up to +10.70% in unseen env.).

****

Vox-E: Text-Guided Voxel Editing of 3D Objects

Etai Sella, Gal Fiebelman, Peter Hedman, Hadar Averbuch-Elor; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 430-440

Large scale text-guided diffusion models have garnered significant attention due to their ability to synthesize diverse images that convey complex visual concepts.

This generative power has more recently been leveraged to perform text-to-3D synthesis. In this work, we present a technique that harnesses the power of latent diffusion models for editing existing 3D objects. Our method takes oriented 2D images of a 3D object as input and learns a grid-based volumetric representation of it. To guide the volumetric representation to conform to a target text prompt, we follow unconditional text-to-3D methods and optimize a Score Distillation Sampling (SDS) loss. However, we observe that combining this diffusion-guided loss with an image-based regularization loss that encourages the representation not to deviate too strongly from the input object is challenging, as it requires achieving two conflicting goals while viewing only structure-and-appearance coupled 2D projections. Thus, we introduce a novel volumetric regularization loss that operates directly in 3D space, utilizing the explicit nature of our 3D representation to enforce correlation between the global structure of the original and edited object. Furthermore, we present a technique that optimizes cross-attention volumetric grids to refine the spatial extent of the edits. Extensive experiments and comparisons demonstrate the effectiveness of our approach in creating a myriad of edits which cannot be achieved by prior works. Our code and data will be made publicly available.

****

Inverse Problem Regularization with Hierarchical Variational Autoencoders

Jean Prost, Antoine Houdard, Andrés Almansa, Nicolas Papadakis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22894-22905

In this paper, we propose to regularize ill-posed inverse problems using a deep hierarchical Variational AutoEncoder (HVAE) as an image prior. The proposed method synthesizes the advantages of i) denoiser-based Plug & Play approaches and ii) generative model based approaches to inverse problems. First, we exploit VAE properties to design an efficient algorithm that benefits from convergence guarantees of Plug-and-Play (PnP) methods. Second, our approach is not restricted to specialized datasets and the proposed PnP-HVAE model is able to solve image restoration problems on natural images of any size. Our experiments show that the proposed PnP-HVAE method is competitive with both SOTA denoiser-based PnP approaches, and other SOTA restoration methods based on generative models. The code for this project is available at https://github.com/jprost76/PnP-HVAE.

****

Unpaired Multi-domain Attribute Translation of 3D Facial Shapes with a Square and Symmetric Geometric Map
Zhenfeng Fan, Zhiheng Zhang, Shuang Yang, Chongyang Zhong, Min Cao, Shihong Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20828-20838
While impressive progress has recently been made in image-oriented facial attribute translation, shape-oriented 3D facial attribute translation remains an unsolved issue. This is primarily limited by the lack of 3D generative models and ineffective usage of 3D facial data. We propose a learning framework for 3D facial attribute translation to relieve these limitations. Firstly, we customize a novel geometric map for 3D shape representation and embed it in an end-to-end generative adversarial network. The geometric map represents 3D shapes symmetrically on a square image grid, while preserving the neighboring relationship of 3D vertices in a local least-square sense. This enables effective learning for the latent representation of data with different attributes. Secondly, we employ a unified and unpaired learning framework for multi-domain attribute translation. It not only makes effective usage of data correlation from multiple domains, but also mitigates the constraint for hardly accessible paired data. Finally, we propose a hierarchical architecture for the discriminator to guarantee robust results against both global and local artifacts. We conduct extensive experiments to demonstrate the advantage of the proposed framework over the state-of-the-art in generating high-fidelity facial shapes. Given an input 3D facial shape, the proposed framework is able to synthesize novel shapes of different attributes, which covers some downstream applications, such as expression transfer, gender translation, and aging. Code at https://github.com/NaughtyZZ/3D_facial_shape_attribute_translation_ssgmap.
****
Passive Ultra-Wideband Single-Photon Imaging
Mian Wei, Sotiris Nousias, Rahul Gulve, David B. Lindell, Kiriakos N. Kutulakos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8135-8146
We consider the problem of imaging a dynamic scene over an extreme range of time scales simultaneously--seconds to picoseconds--and doing so passively, without much light, and without any timing signals from the light source(s) emitting it. Because existing flux estimation techniques for single-photon cameras break down in this regime, we develop a flux probing theory that draws insights from stochastic calculus to enable reconstruction of a pixel's time-varying flux from a stream of monotonically-increasing photon detection timestamps. We use this theory to (1) show that passive free-running SPAD cameras have an attainable frequency bandwidth that spans the entire DC-to-31 GHz range in low-flux conditions, (2) derive a novel Fourier-domain flux reconstruction algorithm that scans this range for frequencies with statistically-significant support in the timestamp data, and (3) ensure the algorithm's noise model remains valid even for very low photon counts or non-negligible dead times. We show the potential of this asynchronous imaging regime by experimentally demonstrating several never-seen-before abilities: (1) imaging a scene illuminated simultaneously by sources operating at vastly different speeds without synchronization (bulbs, projectors, multiple pulsed lasers), (2) passive non-line-of-sight video acquisition, and (3) recording ultra-wideband video, which can be played back later at 30 Hz to show everyday motions--but can also be played a billion times slower to show the propagation of light itself.
****
Template Inversion Attack against Face Recognition Systems using 3D Face Reconstruction
Hatef Otroshi Shahreza, Sébastien Marcel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19662-19672
Face recognition systems are increasingly being used in different applications. In such systems, some features (also known as embeddings or templates) are extracted from each face image. Then, the extracted templates are stored in the system's database during the enrollment stage and are later used for recognition. In

this paper, we focus on template inversion attacks against face recognition systems and introduce a novel method (dubbed GaFaR) to reconstruct 3D face from facial templates. To this end, we use a geometry-aware generator network based on generative neural radiance fields (GNeRF), and learn a mapping from facial templates to the intermediate latent space of the generator network. We train our network with a semi-supervised learning approach using real and synthetic images simultaneously. For the real training data, we use a Generative Adversarial Network (GAN) based framework to learn the distribution of the latent space. For the synthetic training data, where we have the true latent code, we directly train in the latent space of the generator network. In addition, during the inference stage, we also propose optimization on the camera parameters to generate face images to improve the success attack rate (up to 17.14% in our experiments). We evaluate the performance of our method in the whitebox and blackbox attacks against state-of-the-art face recognition models on the LFW and MOBIO datasets. To our knowledge, this paper is the first work on 3D face reconstruction from facial templates. The project page is available at: https://www.idiap.ch/paper/gafar
****
ETran: Energy-Based Transferability Estimation
Mohsen Gholami, Mohammad Akbari, Xinglu Wang, Behnam Kamranian, Yong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18613-18622
This paper addresses the problem of ranking pre-trained models for object detection and image classification. Selecting the best pre-trained model by fine-tuning is an expensive and time-consuming task. Previous works have proposed transferability estimation based on features extracted by the pre-trained models. We argue that quantifying whether the target dataset is in-distribution (IND) or out-of-distribution (OOD) for the pre-trained model is an important factor in the transferability estimation. To this end, we propose ETran, an energy-based transferability assessment metric, which includes three scores: 1) energy score, 2) classification score, and 3) regression score. We use energy-based models to determine whether the target dataset is OOD or IND for the pre-trained model. In contrast to the prior works, ETran is applicable to a wide range of tasks including classification, regression, and object detection (classification+regression). This is the first work that proposes transferability estimation for object detection task. Our extensive experiments on four benchmarks and two tasks show that ETran outperforms previous works on object detection and classification benchmarks by an average of 21% and 12%, respectively, and achieves SOTA in transferability assessment.
****
Predict to Detect: Prediction-guided 3D Object Detection using Sequential Images
Sanmin Kim, Youngseok Kim, In-Jae Lee, Dongsuk Kum; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18057-18066
Recent camera-based 3D object detection methods have introduced sequential frames to improve the detection performance hoping that multiple frames would mitigate the large depth estimation error. Despite improved detection performance, prior works rely on naive fusion methods (e.g., concatenation) or are limited to static scenes (e.g., temporal stereo), neglecting the importance of the motion cue of objects. These approaches do not fully exploit the potential of sequential images and show limited performance improvements. To address this limitation, we propose a novel 3D object detection model, P2D (Predict to Detect), that integrates a prediction scheme into a detection framework to explicitly extract and leverage motion features. P2D predicts object information in the current frame using solely past frames to learn temporal motion features. We then introduce a novel temporal feature aggregation method that attentively exploits Bird's-Eye-View (BEV) features based on predicted object information, resulting in accurate 3D object detection. Experimental results demonstrate that P2D improves mAP and NDS by 3.0% and 3.7% compared to the sequential image-based baseline, proving that incorporating a prediction scheme can significantly improve detection accuracy.
****
Unilaterally Aggregated Contrastive Learning with Hierarchical Augmentation for

Anomaly Detection
Guodong Wang, Yunhong Wang, Jie Qin, Dongming Zhang, Xiuguo Bao, Di Huang; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
 pp. 6888-6897
Anomaly detection (AD), aiming to find samples that deviate from the training di
stribution, is essential in safety-critical applications. Though recent self-sup
ervised learning based attempts achieve promising results by creating virtual ou
tliers, their training objectives are less faithful to AD which requires a conce
ntrated inlier distribution as well as a dispersive outlier distribution. In thi
s paper, we propose Unilaterally Aggregated Contrastive Learning with Hierarchic
al Augmentation (UniCon-HA), taking into account both the requirements above. Sp
ecifically, we explicitly encourage the concentration of inliers and the dispers
ion of virtual outliers via supervised and unsupervised contrastive losses, resp
ectively. Considering that standard contrastive data augmentation for generating
 positive views may induce outliers, we additionally introduce a soft mechanism
to re-weight each augmented inlier according to its deviation from the inlier di
stribution, to ensure a purified concentration. Moreover, to prompt a higher con
centration, inspired by curriculum learning, we adopt an easy-to-hard hierarchic
al augmentation strategy and perform contrastive aggregation at different depths
 of the network based on the strengths of data augmentation. Our method is evalu
ated under three AD settings including unlabeled one-class, unlabeled multi-clas
s, and labeled multi-class, demonstrating its consistent superiority over other
competitors.
****
Learning Image-Adaptive Codebooks for Class-Agnostic Image Restoration
Kechun Liu, Yitong Jiang, Inchang Choi, Jinwei Gu; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 5373-5383
Recent work of discrete generative priors, in the form of codebooks, has shown e
xciting performance for image reconstruction and restoration, since the discrete
 prior space spanned by the codebooks increases the robustness against diverse i
mage degradations. Nevertheless, these methods require separate training of code
books for different image categories, which limits their use to specific image c
ategories only (e.g. face, architecture, etc.), and fail to handle arbitrary nat
ural images. In this paper, we propose AdaCode for learning image-adaptive codeb
ooks for class-agnostic image restoration. Instead of learning a single codebook
 for all categories of images, we learn a set of basis codebooks. For a given in
put image, AdaCode learns a weight map with which we compute a weighted combinat
ion of these basis codebooks for adaptive image restoration. Intuitively, AdaCod
e is a more flexible and expressive discrete generative prior than previous work
. Experimental results show that AdaCode achieves state-of-the-art performance o
n image reconstruction and restoration tasks, including image super-resolution a
nd inpainting.
****
3D Segmentation of Humans in Point Clouds with Synthetic Data
Ayça Takmaz, Jonas Schult, Irem Kaftan, Mertcan Akçay, Bastian Leibe, Robert Sum
ner, Francis Engelmann, Siyu Tang; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2023, pp. 1292-1304
Segmenting humans in 3D indoor scenes has become increasingly important with the
 rise of human-centered robotics and AR/VR applications. To this end, we propose
 the task of joint 3D human semantic segmentation, instance segmentation and mul
ti-human body-part segmentation. Few works have attempted to directly segment hu
mans in cluttered 3D scenes, which is largely due to the lack of annotated train
ing data of humans interacting with 3D scenes. We address this challenge and pro
pose a framework for generating training data of synthetic humans interacting wi
th real 3D scenes. Furthermore, we propose a novel transformer-based model, Huma
n3D, which is the first end-to-end model for segmenting multiple human instances
 and their body-parts in a unified manner. The key advantage of our synthetic da
ta generation framework is its ability to generate diverse and realistic human-s
cene interactions, with highly accurate ground truth. Our experiments show that
pre-training on synthetic data improves performance on a wide variety of 3D huma

n segmentation tasks. Finally, we demonstrate that Human3D outperforms even task -specific state-of-the-art 3D segmentation methods.
****
Mastering Spatial Graph Prediction of Road Networks
Anagnostidis Sotiris, Aurelien Lucchi, Thomas Hofmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5408-5418
Accurately predicting road networks from satellite images requires a global understanding of the network topology. We propose to capture such high-level information by introducing a graph-based framework that given a partially generated graph, sequentially adds new edges. To deal with misalignment between the model predictions and the intended purpose, and to optimize over complex, non-continuous metrics of interest, we adopt a reinforcement learning (RL) approach that nominates modifications that maximize a cumulative reward. As opposed to standard supervised techniques that tend to be more restricted to commonly used surrogate losses, our framework yields more power and flexibility to encode problem-dependent knowledge. Empirical results on several benchmark datasets demonstrate enhanced performance and increased high-level reasoning about the graph topology when using a tree-based search. We further demonstrate the superiority of our approach in handling examples with substantial occlusion and additionally provide evidence that our predictions better match the statistical properties of the ground dataset.
****
IDiff-Face: Synthetic-based Face Recognition through Fizzy Identity-Conditioned Diffusion Model
Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, Naser Damer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19650-19661
The availability of large-scale authentic face databases has been crucial to the significant advances made in face recognition research over the past decade. However, legal and ethical concerns led to the recent retraction of many of these databases by their creators, raising questions about the continuity of future face recognition research without one of its key resources. Synthetic datasets have emerged as a promising alternative to privacy-sensitive authentic data for face recognition development. However, recent synthetic datasets that are used to train face recognition models suffer either from limitations in intra-class diversity or cross-class (identity) discrimination, leading to less optimal accuracies, far away from the accuracies achieved by models trained on authentic data. This paper targets this issue by proposing IDiff-Face, a novel approach based on conditional latent diffusion models for synthetic identity generation with realistic identity variations for face recognition training. Through extensive evaluations, our proposed synthetic-based face recognition approach pushed the limits of state-of-the-art performances, achieving, for example, 98.00% accuracy on the Labeled Faces in the Wild (LFW) benchmark, far ahead from the recent synthetic-based face recognition solutions with 95.40% and bridging the gap to authentic-based face recognition with 99.82% accuracy.
****
Deep Video Demoireing via Compact Invertible Dyadic Decomposition
Yuhui Quan, Haoran Huang, Shengfeng He, Ruotao Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12677-12686
Removing moire patterns from videos recorded on screens or complex textures is known as video demoireing. It is a challenging task as both structures and textures of an image usually exhibit strong periodic patterns, which thus are easily confused with moire patterns and can be significantly erased in the removal process. By interpreting video demoireing as a multi-frame decomposition problem, we propose a compact invertible dyadic network called CIDNet that progressively decouples latent frames and the moire patterns from an input video sequence. Using a dyadic cross-scale coupling structure with coupling layers tailored for multi-scale processing, CIDNet aims at disentangling the features of image patterns from that of moire patterns at different scales, while retaining all latent image features to facilitate reconstruction. In addition, a compressed form for the ne

twork's output is introduced to reduce computational complexity and alleviate overfitting. The experiments show that CIDNet outperforms existing methods and enjoys the advantages in model size and computational efficiency.
****
Rethinking Multi-Contrast MRI Super-Resolution: Rectangle-Window Cross-Attention Transformer and Arbitrary-Scale Upsampling
Guangyuan Li, Lei Zhao, Jiakai Sun, Zehua Lan, Zhanjie Zhang, Jiafu Chen, Zhijie Lin, Huaizhong Lin, Wei Xing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21230-21240
Recently, several methods have explored the potential of multi-contrast magnetic resonance imaging (MRI) super-resolution (SR) and obtain results superior to single-contrast SR methods. However, existing approaches still have two shortcomings: (1) They can only address fixed integer upsampling scales, such as 2x, 3x, and 4x, which require training and storing the corresponding model separately for each upsampling scale in clinic. (2) They lack direct interaction among different windows as they adopt the square window (e.g., 8x8) transformer network architecture, which results in inadequate modelling of longer-range dependencies. Moreover, the relationship between reference images and target images is not fully mined. To address these issues, we develop a novel network for multi-contrast MRI arbitrary-scale SR, dubbed as McASSR. Specifically, we design a rectangle-window cross-attention transformer to establish longer-range dependencies in MR images without increasing computational complexity and fully use reference information. Besides, we propose the reference-aware implicit attention as an upsampling module, achieving arbitrary-scale super-resolution via implicit neural representation, further fusing supplementary information of the reference image. Extensive and comprehensive experiments on both public and clinical datasets show that our McASSR yields superior performance over SOTA methods, demonstrating its great potential to be applied in clinical practice.
****
Domain Generalization via Rationale Invariance
Liang Chen, Yong Zhang, Yibing Song, Anton van den Hengel, Lingqiao Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1751-1760
This paper offers a new perspective to ease the challenge of domain generalization, which involves maintaining robust results even in unseen environments. Our design focuses on the decision-making process in the final classifier layer. Specifically, we propose treating the element-wise contributions to the final results as the rationale for making a decision and representing the rationale for each sample as a matrix. For a well-generalized model, we suggest the rationale matrices for samples belonging to the same category should be similar, indicating the model relies on domain-invariant clues to make decisions, thereby ensuring robust results. To implement this idea, we introduce a rationale invariance loss as a simple regularization technique, requiring only a few lines of code. Our experiments demonstrate that the proposed approach achieves competitive results across various datasets, despite its simplicity. Code is available at https://github.com/liangchen527/RIDG.
****
ProbVLM: Probabilistic Adapter for Frozen Vison-Language Models
Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, Zeynep Akata; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1899-1910
Large-scale vision-language models (VLMs) like CLIP successfully find correspondences between images and text. Through the standard deterministic mapping process, an image or a text sample is mapped to a single vector in the embedding space. This is problematic: as multiple samples (images or text) can abstract the same concept in the physical world, deterministic embeddings do not reflect the inherent ambiguity in the embedding space. We propose ProbVLM, a probabilistic adapter that estimates probability distributions for the embeddings of pre-trained VLMs via inter/intra-modal alignment in a post-hoc manner without needing large-scale datasets or computing. On four challenging datasets, i.e., COCO, Flickr, CU

B, and Oxford-flowers, we estimate the multi-modal embedding uncertainties for t
wo VLMs, i.e., CLIP and BLIP, quantify the calibration of embedding uncertaintie
s in retrieval tasks and show that ProbVLM outperforms other methods. Furthermor
e, we propose active learning and model selection as two real-world downstream t
asks for VLMs and show that the estimated uncertainty aids both tasks. Lastly, w
e present a novel technique for visualizing the embedding distributions using a
large-scale pre-trained latent diffusion model.
****
Towards Open-Set Test-Time Adaptation Utilizing the Wisdom of Crowds in Entropy
Minimization
Jungsoo Lee, Debasmit Das, Jaegul Choo, Sungha Choi; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2023, pp. 16380-16389
Test-time adaptation (TTA) methods, which generally rely on the model's predicti
ons (e.g., entropy minimization) to adapt the source pretrained model to the unl
abeled target domain, suffer from noisy signals originating from 1) incorrect or
 2) open-set predictions. Long-term stable adaptation is hampered by such noisy
signals, so training models without such error accumulation is crucial for pract
ical TTA. To address these issues, including open-set TTA, we propose a simple y
et effective sample selection method inspired by the following crucial empirical
 finding. While entropy minimization compels the model to increase the probabili
ty of its predicted label (i.e., confidence values), we found that noisy samples
 rather show decreased confidence values. To be more specific, entropy minimizat
ion attempts to raise the confidence values of an individual sample's prediction
, but individual confidence values may rise or fall due to the influence of sign
als from numerous other predictions (i.e., wisdom of crowds). Due to this fact,
noisy signals misaligned with such 'wisdom of crowds', generally found in the co
rrect signals, fail to raise the individual confidence values of wrong samples,
despite attempts to increase them. Based on such findings, we filter out the sam
ples whose confidence values are lower in the adapted model than in the original
 model, as they are likely to be noisy. Our method is widely applicable to exist
ing TTA methods and improves their long-term adaptation performance in both imag
e classification (e.g., 49.4% reduced error rates with TENT) and semantic segmen
tation (e.g., 11.7% gain in mIoU with TENT).
****
Scene Graph Contrastive Learning for Embodied Navigation
Kunal Pratap Singh, Jordi Salvador, Luca Weihs, Aniruddha Kembhavi; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10
884-10894
Training effective embodied AI agents often involves expert imitation, specializ
ed components such as maps, or leveraging additional sensors for depth and local
ization. Another approach is to use neural architectures alongside self-supervis
ed objectives which encourage better representation learning. However, in practi
ce, there are few guarantees that these self-supervised objectives encode task-r
elevant information. We propose the Scene Graph Contrastive (SGC) loss, which us
es scene graphs as training-only supervisory signals. The SGC loss does away wit
h explicit graph decoding and instead uses contrastive learning to align an agen
t's representation with a rich graphical encoding of its environment. The SGC lo
ss is simple to implement and encourages representations that encode objects' se
mantics, relationships, and history. By using the SGC loss, we attain gains on t
hree embodied tasks: Object Navigation, Multi-Object Navigation, and Arm Point N
avigation. Finally, we present studies and analyses which demonstrate the abilit
y of our trained representation to encode semantic cues about the environment.
****
Long-Range Grouping Transformer for Multi-View 3D Reconstruction
Liying Yang, Zhenwei Zhu, Xuxin Lin, Jian Nong, Yanyan Liang; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18257-18
267
Nowadays, transformer networks have demonstrated superior performance in many co
mputer vision tasks. In a multi-view 3D reconstruction algorithm following this
paradigm, self-attention processing has to deal with intricate image tokens incl

uding massive information when facing heavy amounts of view input. The curse of
information content leads to the extreme difficulty of model learning. To allevi
ate this problem, recent methods compress the token number representing each vie
w or discard the attention operations between the tokens from different views. O
bviously, they give a negative impact on performance. Therefore, we propose long
-range grouping attention (LGA) based on the divide-and-conquer principle. Token
s from all views are grouped for separate attention operations. The tokens in ea
ch group are sampled from all views and can provide macro representation for the
 resided view. The richness of feature learning is guaranteed by the diversity a
mong different groups. An effective and efficient encoder can be established whi
ch connects inter-view features using LGA and extract intra-view features using
the standard self-attention layer. Moreover, a novel progressive upsampling deco
der is also designed for voxel generation with relatively high resolution. Hingi
ng on the above, we construct a powerful transformer-based network, called LRGT.
 Experimental results on ShapeNet verify our method achieves SOTA accuracy in mu
lti-view reconstruction. Code is available at https://github.com/LiyingCV/Long-R
ange-Grouping-Transformer.
****
Latent-OFER: Detect, Mask, and Reconstruct with Latent Vectors for Occluded Faci
al Expression Recognition
Isack Lee, Eungi Lee, Seok Bong Yoo; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 1536-1546
Most research on facial expression recognition (FER) is conducted in highly cont
rolled environments, but its performance is often unacceptable when applied to r
eal-world situations. This is because when unexpected objects occlude the face,
the FER network faces difficulties extracting facial features and accurately pre
dicting facial expressions. Therefore, occluded FER (OFER) is a challenging prob
lem. Previous studies on occlusion-aware FER have typically required fully annot
ated facial images for training. However, collecting facial images with various
occlusions and expression annotations is time-consuming and expensive. Latent-OF
ER, the proposed method, can detect occlusions, restore occluded parts of the fa
ce as if they were unoccluded, and recognize them, improving FER accuracy. This
approach involves three steps: First, the vision transformer (ViT)-based occlusi
on patch detector masks the occluded position by training only latent vectors fr
om the unoccluded patches using the support vector data description algorithm. S
econd, the hybrid reconstruction network generates the masking position as a com
plete image using the ViT and convolutional neural network (CNN). Last, the expr
ession-relevant latent vector extractor retrieves and uses expression-related in
formation from all latent vectors by applying a CNN-based class activation map.
This mechanism has a significant advantage in preventing performance degradation
 from occlusion by unseen objects. The experimental results on several databases
 demonstrate the superiority of the proposed method over state-of-the-art method
s.
****
DenseShift: Towards Accurate and Efficient Low-Bit Power-of-Two Quantization
Xinlin Li, Bang Liu, Rui Heng Yang, Vanessa Courville, Chao Xing, Vahid Partovi
Nia; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 17010-17020
Efficiently deploying deep neural networks on low-resource edge devices is chall
enging due to their ever-increasing resource requirements. To address this issue
, researchers have proposed multiplication-free neural networks, such as Power-o
f-Two quantization, or also known as Shift networks, which aim to reduce memory
usage and simplify computation. However, existing low-bit Shift networks are not
 as accurate as their full-precision counterparts, typically suffering from limi
ted weight range encoding schemes and quantization loss.
 In this paper, we propose the DenseShift network, which significantly improves
the accuracy of Shift networks, achieving competitive performance to full-precis
ion networks for vision and speech applications. In addition, we introduce a met
hod to deploy an efficient DenseShift network using non-quantized floating-point
 activations, while obtaining 1.6X speed-up over existing methods.

To achieve this, we demonstrate that zero-weight values in low-bit Shift networks do not contribute to model capacity and negatively impact inference computation. To address this issue, we propose a zero-free shifting mechanism that simplifies inference and increases model capacity. We further propose a sign-scale decomposition design to enhance training efficiency and a low-variance random initialization strategy to improve the model's transfer learning performance. Our extensive experiments on various computer vision and speech tasks demonstrate that DenseShift outperforms existing low-bit multiplication-free networks and achieves competitive performance compared to full-precision networks. Furthermore, our proposed approach exhibits strong transfer learning performance without a drop in accuracy. Our code was released on GitHub.
****
Preparing the Future for Continual Semantic Segmentation
Zihan Lin, Zilei Wang, Yixin Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11910-11920
In this study, we focus on Continual Semantic Segmentation (CSS) and present a novel approach to tackle the issue of existing methods struggling to learn new classes. The primary challenge of CSS is to learn new knowledge while retaining old knowledge, which is commonly known as the rigidity-plasticity dilemma. Existing approaches strive to address this by carefully balancing the learning of new and old classes during training on new data. Differently, this work aims to avoid this dilemma fundamentally rather than handling the difficulties involved in it. Specifically, we reveal that this dilemma mainly arises from the greater fluctuation of knowledge for new classes because they have never been learned before the current step. Additionally, the data available in incremental steps are usually inadequate, which can impede the model's ability to learn discriminative features for both new and old classes. To address these challenges, we introduce a novel concept of pre-learning for future knowledge. Our approach entails optimizing the feature space and output space for unlabeled data, which thus enables the model to acquire knowledge for future classes. With this approach, updating the model for new classes becomes as smooth as for old classes, effectively avoiding the rigidity-plasticity dilemma. We conducted extensive experiments and the results demonstrate a significant improvement in the learning of new classes compared to previous state-of-the-art methods.
****
Efficient Computation Sharing for Multi-Task Visual Scene Understanding
Sara Shoouri, Mingyu Yang, Zichen Fan, Hun-Seok Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17130-17141
Solving multiple visual tasks using individual models can be resource-intensive, while multi-task learning can conserve resources by sharing knowledge across different tasks. Despite the benefits of multi-task learning, such techniques can struggle with balancing the loss for each task, leading to potential performance degradation. We present a novel computation- and parameter-sharing framework that balances efficiency and accuracy to perform multiple visual tasks utilizing individually-trained single-task transformers. Our method is motivated by transfer learning schemes to reduce computational and parameter storage costs while maintaining the desired performance. Our approach involves splitting the tasks into a base task and the other sub-tasks, and sharing a significant portion of activations and parameters/weights between the base and sub-tasks to decrease inter-task redundancies and enhance knowledge sharing. The evaluation conducted on NYUD-v2 and PASCAL-context datasets shows that our method is superior to the state-of-the-art transformer-based multi-task learning techniques with higher accuracy and reduced computational resources. Moreover, our method is extended to video stream inputs, further reducing computational costs by efficiently sharing information across the temporal domain as well as the task domain. Our codes are available at https://github.com/sarashoouri/EfficientMTL.
****
Self-supervised Cross-view Representation Reconstruction for Change Captioning
Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Yan, Qingming Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,

pp. 2805-2815
Change captioning aims to describe the difference between a pair of similar imag
es. Its key challenge is how to learn a stable difference representation under p
seudo changes caused by viewpoint change. In this paper, we address this by prop
osing a self-supervised cross-view representation reconstruction (SCORER) networ
k. Concretely, we first design a multi-head token-wise matching to model relatio
nships between cross-view features from similar/dissimilar images. Then, by maxi
mizing cross-view contrastive alignment of two similar images, SCORER learns two
 view-invariant image representations in a self-supervised way. Based on these,
we reconstruct the representations of unchanged objects by cross-attention, thus
 learning a stable difference representation for caption generation. Further, we
 devise a cross-modal backward reasoning to improve the quality of caption. This
 module reversely models a "hallucination" representation with the caption and "
before" representation. By pushing it closer to the "after" representation, we e
nforce the caption to be informative about the difference in a self-supervised m
anner. Extensive experiments show our method achieves the state-of-the-art resul
ts on four datasets. The code is available at https://github.com/tuyunbin/SCORER
.
****
Unify, Align and Refine: Multi-Level Semantic Alignment for Radiology Report Gen
eration
Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, Yuexian Zou; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
 pp. 2863-2874
Automatic radiology report generation has attracted enormous research interest d
ue to its practical value in reducing the workload of radiologists. However, sim
ultaneously establishing global correspondences between the image (e.g., Chest X
-ray) and its related report and local alignments between image patches and keyw
ords remains challenging. To this end, we propose an Unify, Align and then Refin
e (UAR) approach to learn multi-level cross-modal alignments and introduce three
 novel modules: Latent Space Unifier (LSU), Cross-modal Representation Aligner (
CRA) and Text-to-Image Refiner (TIR). Specifically, LSU unifies multimodal data
into discrete tokens, making it flexible to learn common knowledge among modalit
ies with a shared network. The modality-agnostic CRA learns discriminative featu
res via a set of orthonormal basis and a dual-gate mechanism first and then glob
ally aligns visual and textual representations under a triplet contrastive loss.
 TIR boosts token-level local alignment via calibrating text-to-image attention
with a learnable mask. Additionally, we design a two-stage training procedure to
 make UAR gradually grasp cross-modal alignments at different levels, which imit
ates radiologists' workflow: writing sentence by sentence first and then checkin
g word by word. Extensive experiments and analyses on IU-Xray and MIMIC-CXR benc
hmark datasets demonstrate the superiority of our UAR against varied state-of-th
e-art methods.
****
Synthesizing Diverse Human Motions in 3D Indoor Scenes
Kaifeng Zhao, Yan Zhang, Shaofei Wang, Thabo Beeler, Siyu Tang; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14738-
14749
We present a novel method for populating 3D indoor scenes with virtual humans th
at can navigate in the environment and interact with objects in a realistic mann
er. Existing approaches rely on high-quality training sequences that contain cap
tured human motions and the 3D scenes they interact with. However, such interact
ion data are costly, difficult to capture, and can hardly cover the full range o
f plausible human-scene interactions in complex indoor environments. To address
these challenges, we propose a reinforcement learning-based approach that enable
s virtual humans to navigate in 3D scenes and interact with objects realisticall
y and autonomously, driven by learned motion control policies. The motion contro
l policies employ latent motion action spaces, which correspond to realistic mot
ion primitives and are learned from large-scale motion capture data using a powe
rful generative motion model. For navigation in a 3D environment, we propose a s

cene-aware policy with novel state and reward designs for collision avoidance. Combined with navigation mesh-based path-finding algorithms to generate intermediate waypoints, our approach enables the synthesis of diverse human motions navigating in 3D indoor scenes and avoiding obstacles. To generate fine-grained human-object interactions, we carefully curate interaction goal guidance using a marker-based body representation and leverage features based on the signed distance field (SDF) to encode human-scene proximity relations. Our method can synthesize realistic and diverse human-object interactions (e.g., sitting on a chair and then getting up) even for out-of-distribution test scenarios with different object shapes, orientations, starting body positions, and poses. Experimental results demonstrate that our approach outperforms state-of-the-art human-scene interaction synthesis methods in terms of both motion naturalness and diversity. Code, models, and demonstrative video results are publicly available at: https://zkf1997.github.io/DIMOS.
****
Deep Optics for Video Snapshot Compressive Imaging
Ping Wang, Lishun Wang, Xin Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10646-10656
Video snapshot compressive imaging (SCI) aims to capture a sequence of video frames with only a single shot of a 2D detector, whose backbones rest in optical modulation patterns (also known as masks) and a computational reconstruction algorithm. Advanced deep learning algorithms and mature hardware are putting video SCI into practical applications. Yet, there are two clouds in the sunshine of SCI: i) low dynamic range as a victim of high temporal multiplexing, and ii) existing deep learning algorithms' degradation on real system. To address these challenges, this paper presents a deep optics framework to jointly optimize masks and a reconstruction network. Specifically, we first propose a new type of structural mask to realize motionaware and full-dynamic-range measurement. Considering the motion awareness property in measurement domain, we develop an efficient network for video SCI reconstruction using Transformer to capture long-term temporal dependencies, dubbed Res2former. Moreover, sensor response is introduced into the forward model of video SCI to guarantee end-to-end model training close to real system. Finally, we implement the learned structural masks on a digital micro-mirror device. Experimental results on synthetic and real data validate the effectiveness of the proposed framework. We believe this is a miestone for real-world video SCI. The source code and data are available at https://github.com/pwangcs/DeepOpticsSCI.
****
DDIT: Semantic Scene Completion via Deformable Deep Implicit Templates
Haoang Li, Jinhu Dong, Binghui Wen, Ming Gao, Tianyu Huang, Yun-Hui Liu, Daniel Cremers; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21894-21904
Scene reconstructions are often incomplete due to occlusions and limited viewpoints. There have been efforts to use semantic information for scene completion. However, the completed shapes may be rough and imprecise since respective methods rely on 3D convolution and/or lack effective shape constraints. To overcome these limitations, we propose a semantic scene completion method based on deformable deep implicit templates (DDIT). Specifically, we complete each segmented instance in a scene by deforming a template with a latent code. Such a template is expressed by a deep implicit function in the canonical frame. It abstracts the shape prior of a category, and thus can provide constraints on the overall shape of an instance. Latent code controls the deformation of template to guarantee fine details of an instance. For code prediction, we design a neural network that leverages both intra- and inter-instance information. We also introduce an algorithm to transform instances between the world and canonical frames based on geometric constraints and a hierarchical tree. To further improve accuracy, we jointly optimize the latent code and transformation by enforcing the zero-valued isosurface constraint. In addition, we establish a new dataset to solve different problems of existing datasets. Experiments showed that our DDIT outperforms state-of-the-art approaches.

****
Joint Demosaicing and Deghosting of Time-Varying Exposures for Single-Shot HDR I
maging
Jungwoo Kim, Min H. Kim; Proceedings of the IEEE/CVF International Conference on
 Computer Vision (ICCV), 2023, pp. 12292-12301
The quad-Bayer patterned image sensor has made significant improvements in spati
al resolution over recent years due to advancements in image sensor technology.
This has enabled single-shot high-dynamic-range (HDR) imaging using spatially va
rying multiple exposures. Popular methods for multi-exposure array sensors invol
ve varying the gain of each exposure, but this does not effectively change the p
hotoelectronic energy in each exposure. Consequently, HDR images produced using
gain-based exposure variation may suffer from noise and details being saturated.
 To address this problem, we intend to use time-varying exposures in quad-Bayer
patterned sensors. This approach allows long-exposure pixels to receive more pho
ton energy than short- or middle-exposure pixels, resulting in higher-quality HD
R images. However, time-varying exposures are not ideal for dynamic scenes and r
equire an additional deghosting method. To tackle this issue, we propose a singl
e-shot HDR demosaicing method that takes time-varying multiple exposures as inpu
t and jointly solves both the demosaicing and deghosting problems. Our method us
es a feature-extraction module to handle mosaiced multiple exposures and a multi
scale transformer module to register spatial displacements of multiple exposures
 and colors. We also created a dataset of quad-Bayer sensor input with time-vary
ing exposures and trained our network using this dataset. Results demonstrate th
at our method outperforms baseline HDR reconstruction methods with both syntheti
c and real datasets. With our method, we can achieve high-quality HDR images in
challenging lighting conditions.
****
Scene-Aware Feature Matching
Xiaoyong Lu, Yaping Yan, Tong Wei, Songlin Du; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2023, pp. 3704-3713
Current feature matching methods focus on point-level matching, pursuing better
representation learning of individual features, but lacking further understandin
g of the scene. This results in significant performance degradation when handlin
g challenging scenes such as scenes with large viewpoint and illumination change
s. To tackle this problem, we propose a novel model named SAM, which applies att
entional grouping to guide Scene-Aware feature Matching. SAM handles multi-level
 features, i.e., image tokens and group tokens, with attention layers, and group
s the image tokens with the proposed token grouping module. Our model can be tra
ined by ground-truth matches only and produce reasonable grouping results. With
the sense-aware grouping guidance, SAM is not only more accurate and robust but
also more interpretable than conventional feature matching models. Sufficient ex
periments on various applications, including homography estimation, pose estimat
ion, and image matching, demonstrate that our model achieves state-of-the-art pe
rformance.
****
FDViT: Improve the Hierarchical Architecture of Vision Transformer
Yixing Xu, Chao Li, Dong Li, Xiao Sheng, Fan Jiang, Lu Tian, Ashish Sirasao; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
3, pp. 5950-5960
Despite the fact that transformer-based models have yielded great success in com
puter vision tasks, they suffer from the challenge of high computational costs t
hat limits their use on resource-constrained devices. One major reason is that v
ision transformers have redundant calculations since the self-attention operatio
n generates patches with high similarity at a later stage in the network. Hierar
chical architectures have been proposed for vision transformers to alleviate thi
s challenge. However, by shrinking the spatial dimensions to half of the origina
ls with downsampling layers, the challenge is actually overcompensated, as too m
uch information is lost. In this paper, we propose FDViT to improve the hierarch
ical architecture of the vision transformer by using a flexible downsampling lay
er that is not limited to integer stride to smoothly reduce the sizes of the mid

dle feature maps. Furthermore, a masked auto-encoder architecture is used to facilitate the training of the proposed flexible downsampling layer and produces informative outputs. Experimental results on benchmark datasets demonstrate that the proposed method can reduce computational costs while increasing classification performance and achieving state-of-the-art results. For example, the proposed FDViT-S model achieves a top-1 accuracy of 81.5%, which is 1.7 percent points higher than the ViT-S model and reduces 39% FLOPs.
****

Tuning Pre-trained Model via Moment Probing
Mingze Gao, Qilong Wang, Zhenyi Lin, Pengfei Zhu, Qinghua Hu, Jingbo Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11803-11813
Recently, efficient fine-tuning of large-scale pre-trained models has attracted increasing research interests, where linear probing (LP) as a fundamental module is involved in exploiting the final representations for task-dependent classification. However, most of the existing methods focus on how to effectively introduce a few of learnable parameters, and little work pays attention to the commonly used LP module. In this paper, we propose a novel Moment Probing (MP) method to further explore the potential of LP. Distinguished from LP which builds a linear classification head based on the mean of final features (e.g., word tokens for ViT) or classification tokens, our MP performs a linear classifier on feature distribution, which provides a stronger representation ability by exploiting richer statistical information inherent in features. Specifically, we represent feature distribution by its characteristic function, which is efficiently approximated by using first- and second-order moments of features. Furthermore, we propose a multi-head convolutional cross-covariance to compute second-order moments in an efficient and effective manner. By considering that MP could affect feature learning, we introduce a partially shared module to learn two recalibrating parameters (PSRP) for backbones based on MP, namely MP+. Extensive experiments on ten benchmarks using various models show that our MP significantly outperforms LP and is competitive with counterparts at less training cost, while our MP+ achieves state-of-the-art performance.
****

Attention Where It Matters: Rethinking Visual Document Understanding with Selective Region Concentration
Haoyu Cao, Changcun Bao, Chaohu Liu, Huang Chen, Kun Yin, Hao Liu, Yinsong Liu, Deqiang Jiang, Xing Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19517-19527
We propose a novel end-to-end document understanding model called SeRum (SElective Region Understanding Model) for extracting meaningful information from document images, including document analysis, retrieval, and office automation. Unlike state-of-the-art approaches that rely on multi-stage technical schemes and are computationally expensive, SeRum converts document image understanding and recognition tasks into a local decoding process of the vision tokens of interest, using a content-aware token merge module. This mechanism enables the model to pay more attention to regions of interest generated by the query decoder, improving the model's effectiveness and speeding up the decoding speed of the generative scheme. We also designed several pre-training tasks to enhance the understanding and local awareness of the model. Experimental results demonstrate that SeRum achieves state-of-the-art performance on document understanding tasks and competitive results on text spotting tasks. SeRum represents a substantial advancement towards enabling efficient and effective end-to-end document understanding.
****

Task Agnostic Restoration of Natural Video Dynamics
Muhammad Kashif Ali, Dongjin Kim, Tae Hyun Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13534-13544
In many video restoration/translation tasks, image processing operations are naively extended to the video domain by processing each frame independently, disregarding the temporal connection of the video frames. This disregard for the temporal connection often leads to severe temporal inconsistencies. State-Of-The-Art

(SOTA) techniques that address these inconsistencies rely on the availability of unprocessed videos to implicitly siphon and utilize consistent video dynamics to restore the temporal consistency of frame-wise processed videos which often jeopardizes the translation effect. We propose a general framework for this task that learns to infer and utilize consistent motion dynamics from inconsistent videos to mitigate the temporal flicker while preserving the perceptual quality for both the temporally neighboring and relatively distant frames without requiring the raw videos at test time. The proposed framework produces SOTA results on two benchmark datasets, DAVIS and videvo.net, processed by numerous image processing applications. The code and the trained models will be open-sourced upon acceptance.
****
TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis
Mathis Petrovich, Michael J. Black, Gül Varol; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9488-9497
In this paper, we present TMR, a simple yet effective approach for text to 3D human motion retrieval. While previous work has only treated retrieval as a proxy evaluation metric, we tackle it as a standalone task.
 Our method extends the state-of-the-art text-to-motion synthesis model TEMOS, and incorporates a contrastive loss to better structure the cross-modal latent space. We show that maintaining the motion generation loss, along with the contrastive training, is crucial to obtain good performance. We introduce a benchmark for evaluation and provide an in-depth analysis by reporting results on several protocols. Our extensive experiments on the KIT-ML and HumanML3D datasets show that TMR outperforms the prior work by a significant margin, for example reducing the median rank from 54 to 19. Finally, we showcase the potential of our approach on moment retrieval. Our code and models are publicly available at https://mathis.petrovich.fr/tmr.
****
3D Neural Embedding Likelihood: Probabilistic Inverse Graphics for Robust 6D Pose Estimation
Guangyao Zhou, Nishad Gothoskar, Lirui Wang, Joshua B. Tenenbaum, Dan Gutfreund, Miguel Lázaro-Gredilla, Dileep George, Vikash K. Mansinghka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21625-21636
The ability to perceive and understand 3D scenes is crucial for many applications in computer vision and robotics. Inverse graphics is an appealing approach to 3D scene understanding that aims to infer the 3D scene structure from 2D images. In this paper, we introduce probabilistic modeling to the inverse graphics framework to quantify uncertainty and achieve robustness in 6D pose estimation tasks. Specifically, we propose 3D Neural Embedding Likelihood (3DNEL) as a unified probabilistic model over RGB-D images, and develop efficient inference procedures on 3D scene descriptions. 3DNEL effectively combines learned neural embeddings from RGB with depth information to improve robustness in sim-to-real 6D object pose estimation from RGB-D images. Performance on the YCB-Video dataset is on par with state-of-the-art yet is much more robust in challenging regimes. In contrast to discriminative approaches, 3DNEL's probabilistic generative formulation jointly models multiple objects in a scene, quantifies uncertainty in a principled way, and handles object pose tracking under heavy occlusion. Finally, 3DNEL provides a principled framework for incorporating prior knowledge about the scene and objects, which allows natural extension to additional tasks like camera pose tracking from video.
****
Towards Robust Model Watermark via Reducing Parametric Vulnerability
Guanhao Gan, Yiming Li, Dongxian Wu, Shu-Tao Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4751-4761
Deep neural networks are valuable assets considering their commercial benefits and huge demands for costly annotation and computation resources. To protect the copyright of DNNs, backdoor-based ownership verification becomes popular recently, in which the model owner can watermark the model by embedding a specific back

door behavior before releasing it. The defenders (usually the model owners) can identify whether a suspicious third-party model is "stolen" from them based on the presence of the behavior. Unfortunately, these watermarks are proven to be vulnerable to removal attacks even like fine-tuning. To further explore this vulnerability, we investigate the parametric space and find there exist many watermark-removed models in the vicinity of the watermarked one, which may be easily used by removal attacks. Inspired by this finding, we propose a minimax formulation to find these watermark-removed models and recover their watermark behavior. Extensive experiments demonstrate that our method improves the robustness of the model watermarking against parametric changes and numerous watermark-removal attacks. The codes for reproducing our main experiments are available at https://github.com/GuanhaoGan/robust-model-watermarking.
****
SupFusion: Supervised LiDAR-Camera Fusion for 3D Object Detection
Yiran Qin, Chaoqun Wang, Zijian Kang, Ningning Ma, Zhen Li, Ruimao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22014-22024
LiDAR-Camera fusion-based 3D detection is a critical task for automatic driving. In recent years, many LiDAR-Camera fusion approaches sprung up and gained promising performances compared with single-modal detectors, but always lack carefully designed and effective supervision for the fusion process. In this paper, we propose a novel training strategy called SupFusion, which provides an auxiliary feature level supervision for effective LiDAR-Camera fusion and significantly boosts detection performance. Our strategy involves a data enhancement method named Polar Sampling, which densifies sparse objects and trains an assistant model to generate high-quality features as the supervision. These features are then used to train the LiDAR-Camera fusion model, where the fusion feature is optimized to simulate the generated high-quality features. Furthermore, we propose a simple yet effective deep fusion module, which contiguously gains superior performance compared with previous fusion methods with SupFusion strategy. In such a manner, our proposal shares the following advantages. Firstly, SupFusion introduces auxiliary feature-level supervision which could boost LiDAR-Camera detection performance without introducing extra inference costs. Secondly, the proposed deep fusion could continuously improve the detector's abilities. Our proposed SupFusion and deep fusion module is plug-and-play, we make extensive experiments to demonstrate its effectiveness. Specifically, we gain around 2% 3D mAP improvements on KITTI benchmark based on multiple LiDAR-Camera 3D detectors. Our code is available at https://github.com/IranQin/SupFusion.
****
EMMN: Emotional Motion Memory Network for Audio-driven Emotional Talking Face Generation
Shuai Tan, Bin Ji, Ye Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22146-22156
Synthesizing expression is essential to create realistic talking faces. Previous works consider expressions and mouth shapes as a whole and predict them solely from audio inputs. However, the limited information contained in audio, such as phonemes and coarse emotion embedding, may not be suitable as the source of elaborate expressions. Besides, since expressions are tightly coupled to lip motions, generating expression from other sources is tricky and always neglects expression performed on mouth region, leading to inconsistency between them. To tackle the issues, this paper proposes Emotional Motion Memory Net (EMMN) that synthesizes expression overall on the talking face via emotion embedding and lip motion instead of the sole audio. Specifically, we extract emotion embedding from audio and design Motion Reconstruction module to decompose ground truth videos into mouth features and expression features before training, where the latter encode all facial factors about expression. During training, the emotion embedding and mouth features are used as keys, and the corresponding expression features are used as values to create key-value pairs stored in the proposed Motion Memory Net. Hence, once the audio-relevant mouth features and emotion embedding are individually predicted from audio at inference time, we treat them as a query to retrie

ve the best-matching expression features, performing expression overall on the f ace and thus avoiding inconsistent results. Extensive experiments demonstrate th at our method can generate high-quality talking face videos with accurate lip mo vements and vivid expressions on unseen subjects.
****
Rethinking Vision Transformers for MobileNet Size and Speed
Yanyu Li, Ju Hu, Yang Wen, Georgios Evangelidis, Kamyar Salahi, Yanzhi Wang, Ser gey Tulyakov, Jian Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16889-16900
With the success of Vision Transformers (ViTs) in computer vision tasks, recent arts try to optimize the performance and complexity of ViTs to enable efficient deployment on mobile devices. Multiple approaches are proposed to accelerate att ention mechanism, improve inefficient designs, or incorporate mobile-friendly li ghtweight convolutions to form hybrid architectures. However, ViT and its varian ts still have higher latency or considerably more parameters than lightweight CN Ns, even true for the years-old MobileNet. In practice, latency and size are bot h crucial for efficient deployment on resource-constraint hardware. In this work , we investigate a central question, can transformer models run as fast as Mobil eNet and maintain a similar size? We revisit the design choices of ViTs and prop ose a novel supernet with low latency and high parameter efficiency. We further introduce a novel fine-grained joint search strategy for transformer models that can find efficient architectures by optimizing latency and number of parameters simultaneously. The proposed models, EfficientFormerV2, achieve 3.5% higher top -1 accuracy than MobileNetV2 on ImageNet-1K with similar latency and parameters. This work demonstrate that properly designed and optimized vision transformers can achieve high performance even with MobileNet-level size and speed.
****
Implicit Identity Representation Conditioned Memory Compensation Network for Tal king Head video Generation
Fa-Ting Hong, Dan Xu; Proceedings of the IEEE/CVF International Conference on Co mputer Vision (ICCV), 2023, pp. 23062-23072
Talking head video generation aims to animate a human face in a still image with dynamic poses and expressions using motion information derived from a target-dr iving video, while maintaining the person's identity in the source image. Howeve r, dramatic and complex motions in the driving video cause ambiguous generation, because the still source image cannot provide sufficient appearance information for occluded regions or delicate expression variations, which produces severe a rtifacts and significantly degrades the generation quality. To tackle this probl em, we propose to learn a global facial representation space, and design a novel implicit identity representation conditioned memory compensation network, coine d as MCNet, for high-fidelity talking head generation. Specifically, we devise a network module to learn a unified spatial facial meta-memory bank from all trai ning samples, which can provide rich facial structure and appearance priors to c ompensate warped source facial features for the generation. Furthermore, we prop ose an effective query mechanism based on implicit identity representations lear ned from the discrete keypoints of the source image. It can greatly facilitate t he retrieval of more correlated information from the memory bank for the compens ation. Extensive experiments demonstrate that MCNet can learn representative and complementary facial memory, and can clearly outperform previous state-of-the-a rt talking head generation methods on VoxCeleb1 and CelebV datasets.
****
SINC: Self-Supervised In-Context Learning for Vision-Language Tasks
Yi-Syuan Chen, Yun-Zhu Song, Cheng Yu Yeo, Bei Liu, Jianlong Fu, Hong-Han Shuai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15430-15442
Large Pre-trained Transformers exhibit an intriguing capacity for in-context lea rning. Without gradient updates, these models can rapidly construct new predicto rs from demonstrations presented in the inputs. Recent works promote this abilit y in the vision-language domain by incorporating visual information into large l anguage models that can already make in-context predictions. However, these meth

ods could inherit issues in the language domain, such as template sensitivity and hallucination. Also, the scale of these language models raises a significant demand for computations, making learning and operating these models resource-intensive. To this end, we raise a question: "How can we enable in-context learning without relying on the intrinsic in-context ability of large language models?". To answer it, we propose a succinct and general framework, Self-supervised IN-Context learning (SINC), that introduces a meta-model to learn on self-supervised prompts consisting of tailored demonstrations. The learned models can be transferred to downstream tasks for making in-context predictions on-the-fly. Extensive experiments show that SINC outperforms gradient-based methods in various vision-language tasks under few-shot settings. Furthermore, the designs of SINC help us investigate the benefits of in-context learning across different tasks, and the analysis further reveals the essential components for the emergence of in-context learning in the vision-language domain.
****
LEA2: A Lightweight Ensemble Adversarial Attack via Non-overlapping Vulnerable Frequency Regions
Yaguan Qian, Shuke He, Chenyu Zhao, Jiaqiang Sha, Wei Wang, Bin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4510-4521
Recent work shows that well-designed adversarial examples can fool deep neural networks (DNNs). Due to their transferability, adversarial examples can also attack target models without extra information, called black-box attacks. However, most existing ensemble attacks depend on numerous substitute models to cover the vulnerable subspace of a target model. In this work, we find three types of models with non-overlapping vulnerable frequency regions, which can cover a large enough vulnerable subspace. Based on this finding, we propose a lightweight ensemble adversarial attack named LEA2, integrated by standard, weakly robust, and robust models. Moreover, we analyze Gaussian noise from the perspective of frequency and find that Gaussian noise is located in the vulnerable frequency regions of standard models. Therefore, we substitute standard models with Gaussian noise to ensure the use of high-frequency vulnerable regions while reducing attack time consumption. Experiments on several image datasets indicate that LEA^2 achieves better transferability under different defended models compared with extensive baselines and state-of-the-art attacks.
****
Chupa: Carving 3D Clothed Humans from Skinned Shape Priors using 2D Diffusion Probabilistic Models
Byungjun Kim, Patrick Kwon, Kwangho Lee, Myunggi Lee, Sookwan Han, Daesik Kim, Hanbyul Joo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15965-15976
We propose a 3D generation pipeline that uses diffusion models to generate realistic human digital avatars. Due to the wide variety of human identities, poses, and stochastic details, the generation of 3D human meshes has been a challenging problem. To address this, we decompose the problem into 2D normal map generation and normal map-based 3D reconstruction. Specifically, we first simultaneously generate realistic normal maps for the front and backside of a clothed human, dubbed dual normal maps, using a pose-conditional diffusion model. For 3D reconstruction, we "carve" the prior SMPL-X mesh to a detailed 3D mesh according to the normal maps through mesh optimization. To further enhance the high-frequency details, we present a diffusion resampling scheme on both body and facial regions, thus encouraging the generation of realistic digital avatars. We also seamlessly incorporate a recent text-to-image diffusion model to support text-based human identity control. Our method, namely, Chupa, is capable of generating realistic 3D clothed humans with better perceptual quality and identity variety.
****
Unsupervised Domain Adaptive Detection with Network Stability Analysis
Wenzhang Zhou, Heng Fan, Tiejian Luo, Libo Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6986-6995
Domain adaptive detection aims to improve the generality of a detector, learned

from the labeled source domain, on the unlabeled target domain. In this work, drawing inspiration from the concept of stability from the control theory that a robust system requires to remain consistent both externally and internally regardless of disturbances, we propose a novel framework that achieves unsupervised domain adaptive detection through stability analysis. In specific, we treat discrepancies between images and regions from different domains as disturbances, and introduce a novel simple but effective Network Stability Analysis (NSA) framework that considers various disturbances for domain adaptation. Particularly, we explore three types of perturbations including heavy and light image-level disturbances and instance-level disturbance. For each type, NSA performs external consistency analysis on the outputs from raw and perturbed images and/or internal consistency analysis on their features, using teacher-student models. By integrating NSA into Faster R-CNN, we immediately achieve state-of-the-art results. In particular, we set a new record of 52.7% mAP on Cityscapes-to-FoggyCityscapes, showing the potential of NSA for domain adaptive detection. It is worth noticing, our NSA is designed for general purpose, and thus applicable to one-stage detection model (e.g., FCOS) besides the adopted one, as shown by experiments. Code is released at https://github.com/tiankongzhang/NSA.
****
Learning a Room with the Occ-SDF Hybrid: Signed Distance Function Mingled with Occupancy Aids Scene Representation
Xiaoyang Lyu, Peng Dai, Zizhang Li, Dongyu Yan, Yi Lin, Yifan Peng, Xiaojuan Qi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8940-8950
Implicit neural rendering, using signed distance function (SDF) representation with geometric priors like depth or surface normal, has made impressive strides in the surface reconstruction of large-scale scenes. However, applying this method to reconstruct a room-level scene from images may miss structures in low-intensity areas and/or small, thin objects. We have conducted experiments on three datasets to identify limitations of the original color rendering loss and priors-embedded SDF scene representation.Our findings show that the color rendering loss creates an optimization bias against low-intensity areas, resulting in gradient vanishing and leaving these areas unoptimized. To address this issue, we propose a feature-based color rendering loss that utilizes non-zero feature values to bring back optimization signals. Additionally, the SDF representation can be influenced by objects along a ray path, disrupting the monotonic change of SDF values when a single object is present. Accordingly, we explore using the occupancy representation, which encodes each point separately and is unaffected by objects along a querying ray. Our experimental results demonstrate that the joint forces of the feature-based rendering loss and Occ-SDF hybrid representation scheme can provide high-quality reconstruction results, especially in challenging room-level scenarios. The code is available at https://github.com/shawLyu/Occ-SDF_Hybrid.
****
Cloth2Body: Generating 3D Human Body Mesh from 2D Clothing
Lu Dai, Liqian Ma, Shenhan Qian, Hao Liu, Ziwei Liu, Hui Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15007-15017
In this paper, we define and study a new Cloth2Body problem which has a goal of generating 3d human body meshes from a 2D clothing image. Unlike the existing human mesh recovery problem, Cloth2Body needs to address new and emerging challenges raised by the partial observation of the input and the high diversity of the output. Indeed, there are three specific challenges. First, how to locate and pose human bodies into the clothes. Second, how to effectively estimate body shapes out of various clothing types. Finally, how to generate diverse and plausible results from a 2D clothing image. To this end, we propose an end-to-end framework that can accurately estimate 3D body mesh parameterized by pose and shape from a 2D clothing image. Along this line, we first utilize Kinematics-aware Pose Estimation to estimate body pose parameters. 3D skeleton is employed as a proxy followed by an inverse kinematics module to boost the estimation accuracy. We addi

tionally design an adaptive depth trick to align the re-projected 3D mesh better with 2D clothing image by disentangling the effects of object size and camera extrinsic. Next, we propose Physics-informed Shape Estimation to estimate body shape parameters. 3D shape parameters are predicted based on partial body measurements estimated from RGB image, which not only improves pixel-wise human-cloth alignment, but also enables flexible user editing. Finally, we design Evolution based pose generation method , a skeleton transplanting method inspired by genetic algorithms to generate diverse reasonable poses during inference.
As shown by experimental results on both synthetic and real-world data, the proposed framework achieves state-of-the-art performance and can effectively recover natural and diverse 3D body meshes from 2D images that align well with clothing.
****
Spatially and Spectrally Consistent Deep Functional Maps
Mingze Sun, Shiwei Mao, Puhua Jiang, Maks Ovsjanikov, Ruqi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14497-14507
Cycle consistency has long been exploited as a powerful prior for jointly optimizing maps within a collection of shapes. In this paper, we investigate its utility in the approaches of Deep Functional Maps, which are considered state-of-the-art in non-rigid shape matching. We first justify that under certain conditions, the learned maps, when represented in the spectral domain, are already cycle consistent. Furthermore, we identify the discrepancy that spectrally consistent maps are not necessarily spatially, or point-wise, consistent. In light of this, we present a novel design of unsupervised Deep Functional Maps, which effectively enforces the harmony of learned maps under the spectral and the point-wise representation. By taking advantage of cycle consistency, our framework produces state-of-the-art results in mapping shapes even under significant distortions. Beyond that, by independently estimating maps in both spectral and spatial domains, our method naturally alleviates over-fitting in network training, yielding superior generalization performance and accuracy within an array of challenging tests for both near-isometric and non-isometric datasets.
****
Sparse Point Guided 3D Lane Detection
Chengtang Yao, Lidong Yu, Yuwei Wu, Yunde Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8363-8372
3D lane detection usually builds a dense correspondence between the front-view space and the BEV space to estimate lane points in the 3D space. 3D lanes only occupy a small ratio of the dense correspondence, while most correspondence belongs to the redundant background. This sparsity phenomenon bottlenecks valuable computation and raises the computation cost of building a high-resolution correspondence for accurate results. In this paper, we propose a sparse point-guided 3D lane detection, focusing on points related to 3D lanes. Our method runs in a coarse-to-fine manner, including coarse-level lane detection and iterative fine-level sparse point refinements. In coarse-level lane detection, we build a dense but efficient correspondence between the front view and BEV space at a very low resolution to compute coarse lanes. Then in fine-level sparse point refinement, we sample sparse points around coarse lanes to extract local features from the high-resolution front-view feature map. The high-resolution local information brought by sparse points refines 3D lanes in the BEV space hierarchically from low resolution to high resolution. The sparse point guides a more effective information flow and greatly promotes the SOTA result by 3 points on the overall F1-score and 6 points on several hard situations while reducing almost half memory cost and speeding up 2 times.
****
Event-based Temporally Dense Optical Flow Estimation with Sequential Learning
Wachirawit Ponghiran, Chamika Mihiranga Liyanagedera, Kaushik Roy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9827-9836
Event cameras provide an advantage over traditional frame-based cameras when cap

turing fast-moving objects without a motion blur. They achieve this by recording changes in light intensity (known as events), thus allowing them to operate at a much higher frequency and making them suitable for capturing motions in a highly dynamic scene. Many recent studies have proposed methods to train neural networks (NNs) for predicting optical flow from events. However, they often rely on a spatio-temporal representation constructed from events over a fixed interval, such as 10Hz used in training on the DSEC dataset. This limitation restricts the flow prediction to the same interval (10Hz) whereas the fast speed of event cameras, which can operate up to 3kHz, has not been effectively utilized. In this work, we show that a temporally dense flow estimation at 100Hz can be achieved by treating the flow estimation as a sequential problem using two different variants of recurrent networks - Long-short term memory (LSTM) and spiking neural network (SNN). First, We utilize the NN model constructed similar to the popular EV-FlowNet but with LSTM layers to demonstrate the efficiency of our training method. The model not only produces 10x more frequent optical flow than the existing ones, but the estimated flows also have 13% lower errors than predictions from the baseline EV-FlowNet. Second, we construct an EV-FlowNet SNN but with leaky integrate and fire neurons to efficiently capture the temporal dynamics. We found that simple inherent recurrent dynamics of SNN lead to significant parameter reduction compared to the LSTM model. In addition, because of its event-driven computation, the spiking model is estimated to consume only 1.5% energy of the LSTM model, highlighting the efficiency of SNN in processing events and the potential for achieving temporally dense flow.
****

Going Beyond Nouns With Vision & Language Models Using Synthetic Data

Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, Leonid Karlinsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20155-20165

Large-scale pre-trained Vision & Language (VL) models have shown remarkable performance in many applications, enabling replacing a fixed set of supported classes with zero-shot open vocabulary reasoning over (almost arbitrary) natural language prompts. However, recent works have uncovered a fundamental weakness of these models. For example, their difficulty to understand Visual Language Concepts (VLC) that go 'beyond nouns' such as the meaning of non-object words (e.g., attributes, actions, relations, states, etc.), or difficulty in performing compositional reasoning such as understanding the significance of the order of the words in a sentence. In this work, we investigate to which extent purely synthetic data could be leveraged to teach these models to overcome such shortcomings without compromising their zero-shot capabilities. We contribute Synthetic Visual Concepts (SyViC) - a million-scale synthetic dataset and data generation codebase allowing to generate additional suitable data to improve VLC understanding and compositional reasoning of VL models. Additionally, we propose a general VL finetuning strategy for effectively leveraging SyViC towards achieving these improvements. Our extensive experiments and ablations on VL-Checklist, Winoground, and ARO benchmarks demonstrate that it is possible to adapt strong pre-trained VL models with synthetic data significantly enhancing their VLC understanding (e.g. by 9.9% on ARO and 4.3% on VL-Checklist) with under 1% drop in their zero-shot accuracy.
****

Continual Zero-Shot Learning through Semantically Guided Generative Random Walks

Wenxuan Zhang, Paul Janson, Kai Yi, Ivan Skorokhodov, Mohamed Elhoseiny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11574-11585

Learning novel concepts, remembering previous knowledge, and adapting it to future tasks occur simultaneously throughout a human's lifetime. To model such comprehensive abilities, continual zero-shot learning (CZSL) has recently been introduced. However, most existing methods overused the unseen semantic information that may not be continually accessible in realistic settings. In this paper, we address the challenge of continual zero-shot learning where unseen information is not provided during training, by leveraging generative modeling. The heart of the generative-based methods is to learn quality representations from seen classes to improve the generative understanding of the unseen visual space. Motivated by this, we introduce generalization-bound tools and provide the first theoretical explanation for the benefits of generative modeling to CZSL tasks. Guided by the theoretical analysis, we then propose our learning algorithm that employs a novel semantically guided Generative Random Walk (GRW) loss. The GRW loss augments the training by continually encouraging the model to generate realistic and characterized samples to represent the unseen space. Our algorithm achieves state-of-the-art performance on AWA1, AWA2, CUB, and SUN datasets, surpassing existing CZSL methods by 3-7%. The code is available here https://github.com/wx-zhang/IGCZSL

****

Foreground-Background Distribution Modeling Transformer for Visual Object Tracking

Dawei Yang, Jianfeng He, Yinchao Ma, Qianjin Yu, Tianzhu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10117-10127

Visual object tracking is a fundamental research topic with a broad range of applications. Benefiting from the rapid development of Transformer, pure Transformer trackers have achieved great progress. However, the feature learning of these Transformer-based trackers is easily disturbed by complex backgrounds. To address the above limitations, we propose a novel foreground-background distribution modeling transformer for visual object tracking (F-BDMTrack), including a fore-background agent learning (FBAL) module and a distribution-aware attention (DA2) module in a unified transformer architecture. The proposed F-BDMTrack enjoys several merits. First, the proposed FBAL module can effectively mine fore-background information with designed fore-background agents. Second, the DA2 module can suppress the incorrect interaction between foreground and background by modeling fore-background distribution similarities. Finally, F-BDMTrack can extract discriminative features under ever-changing tracking scenarios for more accurate target state estimation. Extensive experiments show that our F-BDMTrack outperforms previous state-of-the-art trackers on eight tracking benchmarks.

****

MeViS: A Large-scale Benchmark for Video Segmentation with Motion Expressions

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2694-2703

This paper strives for motion expressions guided video segmentation, which focuses on segmenting objects in video content based on a sentence describing the motion of the objects. Existing referring video object datasets typically focus on salient objects and use language expressions that contain excessive static attributes that could potentially enable the target object to be identified in a single frame. These datasets downplay the importance of motion in video content for language-guided video object segmentation. To investigate the feasibility of using motion expressions to ground and segment objects in videos, we propose a large-scale dataset called MeViS, which contains numerous motion expressions to indicate target objects in complex environments. We benchmarked 5 existing referring video object segmentation (RVOS) methods and conducted a comprehensive comparison on the MeViS dataset. The results show that current RVOS methods cannot effectively address motion expression-guided video segmentation. We further analyze the challenges and propose a baseline approach for the proposed MeViS dataset. The goal of our benchmark is to provide a platform that enables the development of effective language-guided video segmentation algorithms that leverage motion expressions as a primary cue for object segmentation in complex video scenes. The proposed MeViS dataset has been released at https://henghuiding.github.io/MeViS.

****

OPERA: Omni-Supervised Representation Learning with Hierarchical Supervisions

Chengkun Wang, Wenzhao Zheng, Zheng Zhu, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5559-5570

The pretrain-finetune paradigm in modern computer vision facilitates the success of self-supervised learning, which tends to achieve better transferability than supervised learning. However, with the availability of massive labeled data, a natural question emerges: how to train a better model with both self and full supervision signals? In this paper, we propose Omni-suPErvised Representation leArning with hierarchical supervisions (OPERA) as a solution. We provide a unified perspective of supervisions from labeled and unlabeled data and propose a unified framework of fully supervised and self-supervised learning. We extract a set of hierarchical proxy representations for each image and impose self and full supervisions on the corresponding proxy representations. Extensive experiments on both convolutional neural networks and vision transformers demonstrate the superiority of OPERA in image classification, segmentation, and object detection.
****
GPFL: Simultaneously Learning Global and Personalized Feature Information for Personalized Federated Learning
Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, Jian Cao, Haibing Guan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5041-5051
Federated Learning (FL) is popular for its privacy-preserving and collaborative learning capabilities. Recently, personalized FL (pFL) has received attention for its ability to address statistical heterogeneity and achieve personalization in FL. However, from the perspective of feature extraction, most existing pFL methods only focus on extracting global or personalized feature information during local training, which fails to meet the collaborative learning and personalization goals of pFL. To address this, we propose a new pFL method, named GPFL, to simultaneously learn global and personalized feature information on each client. We conduct extensive experiments on six datasets in three statistically heterogeneous settings and show the superiority of GPFL over ten state-of-the-art methods regarding effectiveness, scalability, fairness, stability, and privacy. Besides, GPFL mitigates overfitting and outperforms the baselines by up to 8.99% in accuracy.
****
Zero-Shot Contrastive Loss for Text-Guided Diffusion Image Style Transfer
Serin Yang, Hyunmin Hwang, Jong Chul Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22873-22882
Diffusion models have shown great promise in text-guided image style transfer, but there is a trade-off between style transformation and content preservation due to their stochastic nature. Existing methods require computationally expensive fine-tuning of diffusion models or additional neural network. To address this, here we propose a zero-shot contrastive loss for diffusion models that doesn't require additional fine-tuning or auxiliary networks. By leveraging patch-wise contrastive loss between generated samples and original image embeddings in the pre-trained diffusion model, our method can generate images with the same semantic content as the source image in a zero-shot manner. Our approach outperforms existing methods while preserving content and requiring no additional training, not only for image style transfer but also for image-to-image translation and manipulation. Our experimental results validate the effectiveness of our proposed method.
****
Efficient Region-Aware Neural Radiance Fields for High-Fidelity Talking Portrait Synthesis
Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, Lin Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7568-7578
This paper presents ER-NeRF, a novel conditional Neural Radiance Fields (NeRF) based architecture for talking portrait synthesis that can concurrently achieve fast convergence, real-time rendering, and state-of-the-art performance with small model size. Our idea is to explicitly exploit the unequal contribution of spatial regions to guide talking portrait modeling. Specifically, to improve the accuracy of dynamic head reconstruction, a compact and expressive NeRF-based Tri-Plane Hash Representation is introduced by pruning empty spatial regions with thre

e planar hash encoders. For speech audio, we propose a Region Attention Module to generate region-aware condition feature via an attention mechanism. Different from existing methods that utilize an MLP-based encoder to learn the cross-modal relation implicitly, the attention mechanism builds an explicit connection between audio features and spatial regions to capture the priors of local motions. Moreover, a direct and fast Adaptive Pose Encoding is introduced to optimize the head-torso separation problem by mapping the complex transformation of the head pose into spatial coordinates. Extensive experiments demonstrate that our method renders better high-fidelity and audio-lips synchronized talking portrait videos, with realistic details and high efficiency compared to previous methods.
****

End2End Multi-View Feature Matching with Differentiable Pose Optimization
Barbara Roessle, Matthias Nießner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 477-487
Erroneous feature matches have severe impact on subsequent camera pose estimation and often require additional, time-costly measures, like RANSAC, for outlier rejection. Our method tackles this challenge by addressing feature matching and pose optimization jointly. To this end, we propose a graph attention network to predict image correspondences along with confidence weights. The resulting matches serve as weighted constraints in a differentiable pose estimation. Training feature matching with gradients from pose optimization naturally learns to down-weight outliers and boosts pose estimation on image pairs compared to SuperGlue by 6.7% on ScanNet. At the same time, it reduces the pose estimation time by over 50% and renders RANSAC iterations unnecessary. Moreover, we integrate information from multiple views by spanning the graph across multiple frames to predict the matches all at once. Multi-view matching combined with end-to-end training improves the pose estimation metrics on Matterport3D by 18.5% compared to SuperGlue.
****

Low-Light Image Enhancement with Illumination-Aware Gamma Correction and Complete Image Modelling Network
Yinglong Wang, Zhen Liu, Jianzhuang Liu, Songcen Xu, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13128-13137
This paper presents a novel network structure with illumination-aware gamma correction and complete image modelling to solve the low-light image enhancement problem. Low-light environments usually lead to less informative large-scale dark areas, directly learning deep representations from low-light images is insensitive to recovering normal illumination. We propose to integrate the effectiveness of gamma correction with the strong modelling capacities of deep networks, which enables the correction factor gamma to be learned in a coarse to elaborate manner via adaptively perceiving the deviated illumination. Because exponential operation introduces high computational complexity, we propose to use Taylor Series to approximate gamma correction, accelerating the training and inference speed. Dark areas usually occupy large scales in low-light images, common local modelling structures, e.g., CNN, SwinIR, are thus insufficient to recover accurate illumination across whole low-light images. We propose a novel Transformer block to completely simulate the dependencies of all pixels across images via a local-to-global hierarchical attention mechanism, so that dark areas could be inferred by borrowing the information from far informative regions in a highly effective manner. Extensive experiments on several benchmark datasets demonstrate that our approach outperforms state-of-the-art methods.
****

Both Diverse and Realism Matter: Physical Attribute and Style Alignment for Rainy Image Generation
Changfeng Yu, Shiming Chen, Yi Chang, Yibing Song, Luxin Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12387-12397
Although considerable progress has been made in the deraining task under synthetic data, it is still a tough problem under real rain scenes, due to the domain g

ap between the synthetic and real data. Besides, difficulties in collecting and labeling diverse real rain images hinder the progress of this field. Consequently, we attempt to promote real rain removal from rain image generation (RIG) perspective. Existing RIG methods mainly focus on diversity but miss realistic, or the realistic but neglect diversity of the generation. To solve this dilemma, we propose a physical alignment and controllable generation network (PCGNet) for diverse and realistic rain generation. Our key idea is to simultaneously utilize the controllability of attributes from synthetic and the realism of appearance from real data. Specifically, we devise a unified framework to disentangle background, rain attributes, and appearance style from synthetic and real data. Then we collaboratively align the factors with a novel semi-supervised weight moving strategy for attribute, an explicit distribution modeling method for real rain style. Furthermore, we pack these aligned factors into the generation model, achieving physical controllable mapping from the attributes to real rainy with image-level and attribute-level consistency loss. Extensive experiments show that PCGNet can effectively generate appealing rainy results, which sifnicantltly improve the performance under synthetic and real scenes for all existing deraining methods.
****
Exploring the Benefits of Visual Prompting in Differential Privacy
Yizhe Li, Yu-Lin Tsai, Chia-Mu Yu, Pin-Yu Chen, Xuebin Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5158-5167
Visual Prompting (VP) is an emerging and powerful technique that allows sample-efficient adaptation to downstream tasks by engineering a well-trained frozen source model. In this work, we explore the benefits of VP in constructing compelling neural network classifiers with differential privacy (DP). We explore and integrate VP into canonical DP training methods and demonstrate its simplicity and efficiency. In particular, we discover that VP in tandem with PATE, a state-of-the-art DP training method that leverages the knowledge transfer from an ensemble of teachers, achieves the state-of-the-art privacy-utility trade-off with minimum expenditure of privacy budget. Moreover, we conduct additional experiments on cross-domain image classification with a sufficient domain gap to further unveil the advantage of VP in DP. Lastly, we also conduct extensive ablation studies to validate the effectiveness and contribution of VP under DP consideration.
****
Single Image Reflection Separation via Component Synergy
Qiming Hu, Xiaojie Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13138-13147
The reflection superposition phenomenon is complex and widely distributed in the real world, which derives various simplified linear and nonlinear formulations of the problem. In this paper, based on the investigation of the weaknesses of existing models, we propose a more general form of the superposition model by introducing a learnable residue term, which can effectively capture residual information during decomposition, guiding the separated layers to be complete. In order to fully capitalize on its advantages, we further design the network structure elaborately, including a novel dual-stream interaction mechanism and a powerful decomposition network with a semantic pyramid encoder. Extensive experiments and ablation studies are conducted to verify our superiority over state-of-the-art approaches on multiple real-world benchmark datasets.
****
Mining bias-target Alignment from Voronoi Cells
Rémi Nahon, Van-Tam Nguyen, Enzo Tartaglione; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4946-4955
Despite significant research efforts, deep neural networks remain vulnerable to biases: this raises concerns about their fairness and limits their generalization. In this paper, we propose a bias-agnostic approach to mitigate the impact of biases in deep neural networks. Unlike traditional debiasing approaches, we rely on a metric to quantify "bias alignment/misalignment" on target classes and use this information to discourage the propagation of bias-target alignment information through the network. We conduct experiments on several commonly used datase

ts for debiasing and compare our method with supervised and bias-specific
 approaches. Our results indicate that the proposed method achieves comparable p
erformance to state-of-the-art supervised approaches, despite being bias-agnosti
c, even in the presence of multiple biases in the same sample.
****
The Victim and The Beneficiary: Exploiting a Poisoned Model to Train a Clean Mod
el on Poisoned Data
Zixuan Zhu, Rui Wang, Cong Zou, Lihua Jing; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 155-164
Recently, backdoor attacks have posed a serious security threat to the training
process of deep neural networks (DNNs). The attacked model behaves normally on b
enign samples but outputs a specific result when the trigger is present. However
, compared with the rocketing progress of backdoor attacks, existing defenses ar
e difficult to deal with these threats effectively or require benign samples to
work, which may be unavailable in real scenarios. In this paper, we find that th
e poisoned samples and benign samples can be distinguished with prediction entro
py. This inspires us to propose a novel dual-network training framework: The Vic
tim and The Beneficiary (V&B), which exploits a poisoned model to train a clean
model without extra benign samples. Firstly, we sacrifice the Victim network to
be a powerful poisoned sample detector by training on suspicious samples. Second
ly, we train the Beneficiary network on the credible samples selected by the Vic
tim to inhibit backdoor injection. Thirdly, a semi-supervised suppression strate
gy is adopted for erasing potential backdoors and improving model performance. F
urthermore, to better inhibit missed poisoned samples, we propose a strong data
augmentation method, AttentionMix, which works well with our proposed V&B framew
ork. Extensive experiments on two widely used datasets against 6 state-of-the-ar
t attacks demonstrate that our framework is effective in preventing backdoor inj
ection and robust to various attacks while maintaining the performance on benign
 samples. Our code is available at https://github.com/Zixuan-Zhu/VaB.
****
DIFFGUARD: Semantic Mismatch-Guided Out-of-Distribution Detection Using Pre-Trai
ned Diffusion Models
Ruiyuan Gao, Chenchen Zhao, Lanqing Hong, Qiang Xu; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 1579-1589
Given a classifier, the inherent property of semantic Out-of-Distribution (OOD)
samples is that their contents differ from all legal classes in terms of semanti
cs, namely semantic mismatch. There is a recent work that directly applies it to
 OOD detection, which employs a conditional Generative Adversarial Network (cGAN
) to enlarge semantic mismatch in the image space. While achieving remarkable OO
D detection performance on small datasets, it is not applicable to ImageNet-scal
e datasets due to the difficulty in training cGANs with both input images and la
bels as conditions.
 As diffusion models are much easier to train and amenable to various conditions
 compared to cGANs, in this work, we propose to directly use pre-trained diffusi
on models for semantic mismatch-guided OOD detection, named DiffGuard. Specifica
lly, given an OOD input image and the predicted label from the classifier, we tr
y to enlarge the semantic difference between the reconstructed OOD image under t
hese conditions and the original input image. We also present several test-time
techniques to further strengthen such differences. Experimental results show tha
t DiffGuard is effective on both Cifar-10 and hard cases of the large-scale Imag
eNet, and it can be easily combined with existing OOD detection techniques to ac
hieve state-of-the-art OOD detection results.
****
Identity-Seeking Self-Supervised Representation Learning for Generalizable Perso
n Re-Identification
Zhaopeng Dou, Zhongdao Wang, Yali Li, Shengjin Wang; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2023, pp. 15847-15858
This paper aims to learn a domain-generalizable (DG) person re-identification (R
eID) representation from large-scale videos without any annotation. Prior DG ReI
D methods employ limited labeled data for training due to the high cost of annot

ation, which restricts further advances. To overcome the barriers of data and annotation, we propose to utilize large-scale unsupervised data for training. The key issue lies in how to mine identity information. To this end, we propose an Identity-seeking Self-supervised Representation learning (ISR) method. ISR constructs positive pairs from inter-frame images by modeling the instance association as a maximum-weight bipartite matching problem. A reliability-guided contrastive loss is further presented to suppress the adverse impact of noisy positive pairs, ensuring that reliable positive pairs dominate the learning process. The training cost of ISR scales approximately linearly with the data size, making it feasible to utilize large-scale data for training. The learned representation exhibits superior generalization ability. Without human annotation and fine-tuning, ISR achieves 87.0% Rank-1 on Market-1501 and 56.4% Rank-1 on MSMT17, outperforming the best supervised domain-generalizable method by 5.0% and 19.5%, respectively. In the pre-training-to-fine-tuning scenario, ISR achieves state-of-the-art performance, with 88.4% Rank-1 on MSMT17.
****
3D-Aware Generative Model for Improved Side-View Image Synthesis
Kyungmin Jo, Wonjoon Jin, Jaegul Choo, Hyunjoon Lee, Sunghyun Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22862-22872
While recent 3D-aware generative models have shown photo-realistic image synthesis with multi-view consistency, the synthesized image quality degrades depending on the camera pose (e.g., a face with a blurry and noisy boundary at a side viewpoint). Such degradation is mainly caused by the difficulty of learning both pose consistency and photo-realism simultaneously from a dataset with heavily imbalanced poses. In this paper, we propose SideGAN, a novel 3D GAN training method to generate photo-realistic images irrespective of the camera pose, especially for faces of side-view angles. To ease the challenging problem of learning photo-realistic and pose-consistent image synthesis, we split the problem into two subproblems, each of which can be solved more easily. Specifically, we formulate the problem as a combination of two simple discrimination problems, one of which learns to discriminate whether a synthesized image looks real or not, and the other learns to discriminate whether a synthesized image agrees with the camera pose. Based on this, we propose a dual-branched discriminator with two discrimination branches. We also propose a pose-matching loss to learn the pose consistency of 3D GANs. In addition, we present a pose sampling strategy to increase learning opportunities for steep angles in a pose-imbalanced dataset. With extensive validation, we demonstrate that our approach enables 3D GANs to generate high-quality geometries and photo-realistic images irrespective of the camera pose.
****
Tracking Anything with Decoupled Video Segmentation
Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, Joon-Young Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1316-1326
Training data for video segmentation are expensive to annotate. This impedes extensions of end-to-end algorithms to new video segmentation tasks, especially in large-vocabulary settings. To 'track anything' without training on video data for every individual task, we develop a decoupled video segmentation approach (DEVA), composed of task-specific image-level segmentation and class/task-agnostic bi-directional temporal propagation. Due to this design, we only need an image-level model for the target task (which is cheaper to train) and a universal temporal propagation model which is trained once and generalizes across tasks. To effectively combine these two modules, we use bi-directional propagation for (semi-)online fusion of segmentation hypotheses from different frames to generate a coherent segmentation. We show that this decoupled formulation compares favorably to end-to-end approaches in several data-scarce tasks including large-vocabulary video panoptic segmentation, open-world video segmentation, referring video segmentation, and unsupervised video object segmentation.
Code is available at: https://hkchengrex.github.io/Tracking-Anything-with-DEVA.
****

Generative Gradient Inversion via Over-Parameterized Networks in Federated Learn ing
Chi Zhang, Zhang Xiaoman, Ekanut Sotthiwat, Yanyu Xu, Ping Liu, Liangli Zhen, Yo ng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5126-5135
Federated learning has gained recognitions as a secure approach for safeguarding local private data in collaborative learning. But the advent of gradient invers ion research has posed significant challenges to this premise by enabling a thir d-party to recover groundtruth images via gradients. While prior research has pr edominantly focused on low-resolution images and small batch sizes, this study h ighlights the feasibility of reconstructing complex images with high resolutions and large batch sizes. The success of the proposed method is contingent on cons tructing an over-parameterized convolutional network, so that images are generat ed before fitting to the gradient matching requirement. Practical experiments de monstrate that the proposed algorithm achieves high-fidelity image recovery, sur passing state-of-the-art competitors that commonly fail in more intricate scenar ios. Consequently, our study shows that local participants in a federated learni ng system are vulnerable to potential data leakage issues. Source code is availa ble at https://github.com/czhang024/CI-Net.
****
EQ-Net: Elastic Quantization Neural Networks
Ke Xu, Lei Han, Ye Tian, Shangshang Yang, Xingyi Zhang; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2023, pp. 1505-1514
Current model quantization methods have shown their promising capability in redu cing storage space and computation complexity. However, due to the diversity of quantization forms supported by different hardware, one limitation of existing s olutions is that usually require repeated optimization for different scenarios. How to construct a model with flexible quantization forms has been less studied. In this paper, we explore a one-shot network quantization regime, named Elastic Quantization Neural Networks (EQ-Net), which aims to train a robust weight-shar ing quantization supernet. First of all, we propose an elastic quantization spac e (including elastic bit-width, granularity, and symmetry) to adapt to various m ainstream quantitative forms. Secondly, we propose the Weight Distribution Regul arization Loss (WDR-Loss) and Group Progressive Guidance Loss (GPG-Loss) to brid ge the inconsistency of the distribution for weights and output logits in the el astic quantization space gap. Lastly, we incorporate genetic algorithms and the proposed Conditional Quantization-Aware Accuracy Predictor (CQAP) as an estimato r to quickly search mixed-precision quantized neural networks in supernet. Exten sive experiments demonstrate that our EQ-Net is close to or even better than its static counterparts as well as state-of-the-art robust bit-width methods. Code can be available at https://github.com/xuke225/EQ-Net.git
****
OxfordTVG-HIC: Can Machine Make Humorous Captions from Images?
Runjia Li, Shuyang Sun, Mohamed Elhoseiny, Philip Torr; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2023, pp. 20293-20303
This paper presents OxfordTVG-HIC (Humorous Image Captions), a large-scale datas et for humour generation and understanding. Humour is an abstract, subjective, a nd context-dependent cognitive construct involving several cognitive factors, ma king it a challenging task to generate and interpret. Hence, humour generation a nd understanding can serve as a new task for evaluating the ability of deep-lear ning methods to process abstract and subjective information. Due to the scarcity of data, humour-related generation tasks such as captioning remain underexplore d. To address this gap, OxfordTVG-HIC offers approximately 2.9M image-text pairs with humour scores to train a generalizable humour captioning model. Contrary t o existing captioning datasets, OxfordTVG-HIC features a wide range of emotional and semantic diversity resulting in out-of-context examples that are particular ly conducive to generating humour. Moreover, OxfordTVG-HIC is curated devoid of offensive content. We also show how OxfordTVGHIC can be leveraged for evaluating the humour of a generated text. Through explainability analysis of the trained models, we identify the visual and linguistic cues influential for evoking humou

r prediction (and generation). We observe qualitatively that these cues are aligned with the benign violation theory of humour in cognitive psychology.
****
Exploring Open-Vocabulary Semantic Segmentation from CLIP Vision Encoder Distillation Only
Jun Chen, Deyao Zhu, Guocheng Qian, Bernard Ghanem, Zhicheng Yan, Chenchen Zhu, Fanyi Xiao, Sean Chang Culatana, Mohamed Elhoseiny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 699-710
Semantic segmentation is a crucial task in computer vision that involves segmenting images into semantically meaningful regions at the pixel level. However, existing approaches often rely on expensive human annotations as supervision for model training, limiting their scalability to large, unlabeled datasets. To address this challenge, we present ZeroSeg, a novel method that leverages the existing pretrained vision-language (VL) model (e.g. CLIP vision encoder) to train open-vocabulary zero-shot semantic segmentation models. Although acquired extensive knowledge of visual concepts, it is non-trivial to exploit knowledge from these VL models to the task of semantic segmentation, as they are usually trained at an image level. ZeroSeg overcomes this by distilling the visual concepts learned by VL models into a set of segment tokens, each summarizing a localized region of the target image. We evaluate ZeroSeg on multiple popular segmentation benchmarks, including PASCAL VOC 2012, PASCAL Context, and COCO, in a zero-shot manner Our approach achieves state-of-the-art performance when compared to other zero-shot segmentation methods under the same training data, while also performing competitively compared to strongly supervised methods. Finally, we also demonstrated the effectiveness of ZeroSeg on open-vocabulary segmentation, through both human studies and qualitative visualizations. The code is publicly available at https://github.com/facebookresearch/ZeroSeg
****
EDAPS: Enhanced Domain-Adaptive Panoptic Segmentation
Suman Saha, Lukas Hoyer, Anton Obukhov, Dengxin Dai, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19234-19245
With autonomous industries on the rise, domain adaptation of the visual perception stack is an important research direction due to the cost savings promise. Much prior art was dedicated to domain-adaptive semantic segmentation in the synthetic-to-real context. Despite being a crucial output of the perception stack, panoptic segmentation has been largely overlooked by the domain adaptation community. Therefore, we revisit well-performing domain adaptation strategies from other fields, adapt them to panoptic segmentation, and show that they can effectively enhance panoptic domain adaptation. Further, we study the panoptic network design and propose a novel architecture (EDAPS) designed explicitly for domain-adaptive panoptic segmentation. It uses a shared, domain-robust transformer encoder to facilitate the joint adaptation of semantic and instance features, but task-specific decoders tailored for the specific requirements of both domain-adaptive semantic and instance segmentation. As a result, the performance gap seen in challenging panoptic benchmarks is substantially narrowed. EDAPS significantly improves the state-of-the-art performance for panoptic segmentation UDA by a large margin of 20% on SYNTHIA-to-Cityscapes and even 72% on the more challenging SYNTHIA-to-Mapillary Vistas. The implementation is available at https://github.com/susaha/edaps.
****
Parallax-Tolerant Unsupervised Deep Image Stitching
Lang Nie, Chunyu Lin, Kang Liao, Shuaicheng Liu, Yao Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7399-7408
Traditional image stitching approaches tend to leverage increasingly complex geometric features (point, line, edge, etc.) for better performance. However, these hand-crafted features are only suitable for specific natural scenes with adequate geometric structures. In contrast, deep stitching schemes overcome adverse conditions by adaptively learning robust semantic features, but they cannot handle large-parallax cases.

To solve these issues, we propose a parallax-tolerant unsupervised deep image stitching technique. First, we propose a robust and flexible warp to model the image registration from global homography to local thin-plate spline motion. It provides accurate alignment for overlapping regions and shape preservation for non-overlapping regions by joint optimization concerning alignment and distortion. Subsequently, to improve the generalization capability, we design a simple but effective iterative strategy to enhance the warp adaption in cross-dataset and cross-resolution applications. Finally, to further eliminate the parallax artifacts, we propose to composite the stitched image seamlessly by unsupervised learning for seam-driven composition masks. Compared with existing methods, our solution is parallax-tolerant and free from laborious designs of complicated geometric features for specific scenes. Extensive experiments show our superiority over the SoTA methods, both quantitatively and qualitatively. The code will be available soon.
****
Scratch Each Other's Back: Incomplete Multi-Modal Brain Tumor Segmentation via Category Aware Group Self-Support Learning
Yansheng Qiu, Delin Chen, Hongdou Yao, Yongchao Xu, Zheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21317-21326
Although Magnetic Resonance Imaging (MRI) is very helpful for brain tumor segmentation and discovery, it often lacks some modalities in clinical practice. As a result, degradation of prediction performance is inevitable. According to current implementations, different modalities are considered to be independent and non-interfering with each other during the training process of modal feature extraction, however they are complementary. In this paper, considering the sensitivity of different modalities to diverse tumor regions, we propose a Category Aware Group Self-Support Learning framework, called GSS, to make up for the information deficit among the modalities in the individual modal feature extraction phase. Precisely, within each prediction category, predictions of all modalities form a group, where the prediction with the most extraordinary sensitivity is selected as the group leader. Collaborative efforts between group leaders and members identify the communal learning target with high consistency and certainty. As our minor contribution, we introduce a random mask to reduce the possible biases. GSS adopts the standard training strategy without specific architectural choices and thus can be easily plugged into existing incomplete multi-modal brain tumor segmentation. Remarkably, extensive experiments on BraTS2020, BraTS2018, and BraTS2015 datasets demonstrate that GSS can improve the performance of existing SOTA algorithms by 1.27-3.20% in Dice on average. The code is released at https://github.com/qysgithubopen/GSS.
****
SFHarmony: Source Free Domain Adaptation for Distributed Neuroimaging Analysis
Nicola K Dinsdale, Mark Jenkinson, Ana IL Namburete; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11494-11505
To represent the biological variability of clinical neuroimaging populations, it is vital to be able to combine data across scanners and studies. However, different MRI scanners produce images with different characteristics, resulting in a domain shift known as the 'harmonisation problem'. Additionally, neuroimaging data is inherently personal in nature, leading to data privacy concerns when sharing the data. To overcome these barriers, we propose an Unsupervised Source-Free Domain Adaptation (SFDA) method, SFHarmony. Through modelling the imaging features as a Gaussian Mixture Model and minimising an adapted Bhattacharyya distance between the source and target features, we can create a model that performs well for the target data whilst having a shared feature representation across the data domains, without needing access to the source data for adaptation or target labels. We demonstrate the performance of our method on simulated and real domain shifts, showing that the approach is applicable to classification, segmentation and regression tasks, requiring no changes to the algorithm. Our method outperforms existing SFDA approaches across a range of realistic data scenarios, demonstrating the potential utility of our approach for MRI harmonisation and general

SFDA problems. Our code is available at https://github.com/nkdinsdale/SFHarmony.
****
M2T: Masking Transformers Twice for Faster Decoding
Fabian Mentzer, Eirikur Agustson, Michael Tschannen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5340-5349
We show how bidirectional transformers trained for masked token prediction can be applied to neural image compression to achieve state-of-the-art results.
Such models were previously used for image_generation_ by progressive sampling groups of masked tokens according to uncertainty-adaptive schedules.
Unlike these works, we demonstrate that predefined, deterministic schedules perform as well or better for image compression.
This insight allows us to use masked attention during training in addition to masked inputs, and activation caching during inference, to significantly speed up our models (4x higher inference speed) at a small increase in bitrate.
****
CoIn: Contrastive Instance Feature Mining for Outdoor 3D Object Detection with Very Limited Annotations
Qiming Xia, Jinhao Deng, Chenglu Wen, Hai Wu, Shaoshuai Shi, Xin Li, Cheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6254-6263
Recently, 3D object detection with sparse annotations has received great attention. However, current detectors usually perform poorly under very limited annotations. To address this problem, we propose a novel Contrastive Instance feature mining method, named CoIn. To better identify indistinguishable features learned through limited supervision, we design a Multi-Class contrastive learning module (MCcont) to enhance feature discrimination.
Meanwhile, we propose a feature-level pseudo-label mining framework consisting of an instance feature mining module (InF-Mining) and a Labeled-to-Pseudo contrastive learning module (LPcont). These two modules exploit latent instances in feature space to supervise the training of detectors with limited annotations. Extensive experiments with KITTI dataset, Waymo open dataset, and nuScenes dataset show that under limited annotations, our method greatly improves the performance of baseline detectors: CenterPoint, Voxel-RCNN, and CasA. Combining CoIn with an iterative training strategy, we propose a CoIn++ pipeline, which requires only 2% annotations in the KITTI dataset to achieve performance comparable to the fully supervised methods. The code is available at https://github.com/xmuqimingxia/CoIn.
****
3D Human Mesh Recovery with Sequentially Global Rotation Estimation
Dongkai Wang, Shiliang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14953-14962
Model-based 3D human mesh recovery aims to reconstruct a 3D human body mesh by estimating its parameters from monocular RGB images. Most of recent works adopt the Skinned Multi-Person Linear (SMPL) model to regress relative rotations for each body joint along the kinematics chain. This pipeline needs to transform each relative rotation matrix into a global rotation matrix to articulate the canonical mesh, and suffers from accumulated errors along the kinematics chain. This paper proposes to directly estimate the global rotation of each joint to avoid error accumulation and pursue better accuracy. The proposed Sequentially Global Rotation Estimation (SGRE) directly predicts the global rotation matrix of each joint on the kinematics chain. SGRE features a residual learning module to leverage complementary features and previously predicted rotations of parent joints to guide the estimation of subsequent child joints. Thanks to this global estimation pipeline and residual learning module, SGRE alleviates error accumulation and produces more accurate 3D human mesh. It can be flexibly integrated into existing regression-based methods and achieves superior performance on various benchmarks. For example, it improves the latest method 3DCrowdNet by 3.3 mm MPJPE and 5.0 mm PVE on 3DPW dataset and 3.2 AP on COCO dataset, respectively.
****
DREAMWALKER: Mental Planning for Continuous Vision-Language Navigation

Hanqing Wang, Wei Liang, Luc Van Gool, Wenguan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10873-10883
VLN-CE is a recently released embodied task, where AI agents need to navigate a freely traversable environment to reach a distant target location, given language instructions. It poses great challenges due to the huge space of possible strategies. Driven by the belief that the ability to anticipate the consequences of future actions is crucial for the emergence of intelligent and interpretable planning behavior, we propose Dreamwalker --- a world model based VLN-CE agent. The world model is built to summarize the visual, topological, and dynamic properties of the complicated continuous environment into a discrete, structured, and compact representation. Dreamwalker can simulate and evaluate possible plans entirely in such internal abstract world, before executing costly actions. As opposed to existing model-free VLN-CE agents simply making greedy decisions in the real world, which easily results in shortsighted behaviors, Dreamwalker is able to make strategic planning through large amounts of "mental experiments." Moreover, the imagined future scenarios reflect our agent's intention, making its decision-making process more transparent. Extensive experiments and ablation studies on VLN-CE dataset confirm the effectiveness of the proposed approach and outline fruitful directions for future work. Our code will be released.
****
Computation and Data Efficient Backdoor Attacks
Yutong Wu, Xingshuo Han, Han Qiu, Tianwei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4805-4814
Backdoor attacks against deep learning have been widely studied. Various attack techniques have been proposed for different domains and paradigms, e.g., image, point cloud, natural language processing, transfer learning, etc. These works normally adopt the data poisoning strategy to embed the backdoor. They randomly select samples from the benign training set for poisoning, without considering the distinct contribution of each sample to the backdoor effectiveness, making the attack less optimal. A recent work (IJCAI-22) proposed to use the forgetting score to measure the importance of each poisoned sample and then filter out redundant data for effective backdoor training. However, this method is empirically designed without theoretical proofing. It is also very time-consuming as it needs to go through almost all the training stages for data selection. To address such limitations, we propose a novel confidence-based scoring methodology, which can efficiently measure the contribution of each poisoning sample based on the distance posteriors. We further introduce a greedy search algorithm to find the most informative samples for backdoor injection more promptly. Experimental evaluations on both 2D image and 3D point cloud classification tasks show that our approach can achieve comparable performance or even surpass the forgetting score-based searching method while requiring only several extra epochs' computation of a standard training process.
****
Agglomerative Transformer for Human-Object Interaction Detection
Danyang Tu, Wei Sun, Guangtao Zhai, Wei Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21614-21624
We propose an agglomerative Transformer (AGER) that enables Transformer-based human-object interaction (HOI) detectors to flexibly exploit extra instance-level cues in a single-stage and end-to-end manner for the first time. AGER acquires instance tokens by dynamically clustering patch tokens and aligning cluster centres to instances with textual guidance, thus enjoying two benefits: 1) Intergrality: each instance token is encouraged to contain all discriminative feature regions of an instance, which demonstrates a significant improvement in the extraction of different instance-level cues, and subsequently leads to a new state-of-the-art performance of HOI detection with 36.75 mAP on HICO-Det. 2) Efficiency: the dynamical clustering mechanism allows AGER to generate instance tokens jointly with the feature learning of the Transformer encoder, eliminating the need of an additional object detector or instance decoder in prior methods, thus allowing the extraction of desirable extra cues for HOI detection in a single-stage and end-to-end pipeline. Concretely, AGER reduces GFLOPs by 8.5% and improves FPS by

36%, even compared to a vanilla DETR-like pipeline without extra cue extraction
.
****
Decouple Before Interact: Multi-Modal Prompt Learning for Continual Visual Quest
ion Answering
Zi Qian, Xin Wang, Xuguang Duan, Pengda Qin, Yuhong Li, Wenwu Zhu; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 295
3-2962
In the real world, a desirable Visual Question Answering model is expected to pr
ovide correct answers to new questions and images in a continual setting (recogn
ized as CL-VQA). However, existing works formulate CLVQA from a vision-only or l
anguage-only perspective, and straightforwardly apply the uni-modal continual le
arning (CL) strategies to this multi-modal task, which is improper and suboptima
l. On the one hand, such a partial formulation may result in limited evaluations
. On the other hand, neglecting the interactions between modalities will lead to
 poor performance. To tackle these challenging issues, we propose a comprehensiv
e formulation for CL-VQA from the perspective of multi-modal vision-language fus
ion. Based on our formulation, we further propose MulTi-Modal PRompt LearnIng wi
th DecouPLing bEfore InTeraction (TRIPLET), a novel approach that builds on a pr
e-trained vision-language model and consists of decoupled prompts and prompt int
eraction strategies to capture the complex interactions between modalities. In p
articular, decoupled prompts contain learnable parameters that are decoupled w.r
.t different aspects, and the prompt interaction strategies are in charge of mod
eling interactions between inputs and prompts. Additionally, we build two CL-VQA
 benchmarks for a more comprehensive evaluation. Extensive experiments demonstra
te that our TRIPLET outperforms state-of-the-art methods in both uni-modal and m
ulti-modal continual settings for CL-VQA.
****
Rethinking Fast Fourier Convolution in Image Inpainting
Tianyi Chu, Jiafu Chen, Jiakai Sun, Shuobin Lian, Zhizhong Wang, Zhiwen Zuo, Lei
 Zhao, Wei Xing, Dongming Lu; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2023, pp. 23195-23205
Recently proposed image inpainting method LaMa builds its network upon Fast Four
ier Convolution (FFC), which was originally proposed for high-level vision tasks
 like image classification. FFC empowers the fully convolutional network to have
 a global receptive field in its early layers. Thanks to the unique character of
 the FFC module, LaMa has the ability to produce robust repeating texture, which
 can not be achieved by the previous inpainting methods. However, is the vanilla
 FFC module suitable for low-level vision tasks like image inpainting? In this p
aper, we analyze the fundamental flaws of using FFC in image inpainting, which a
re 1) spectrum shifting, 2) unexpected spatial activation, and 3) limited freque
ncy receptive field. Such flaws make FFC-based inpainting framework difficult in
 generating complicated texture and performing faithful reconstruction. Based on
 the above analysis, we propose a novel Unbiased Fast Fourier Convolution (UFFC)
 module, which modifies the vanilla FFC module with 1) range transform and inver
se transform, 2) absolute position embedding, 3) dynamic skip connection, and 4)
 adaptive clip, to overcome such flaws, achieving better inpainting results. Ext
ensive experiments on several benchmark datasets demonstrate the effectiveness o
f our method, outperforming the state-of-the-art methods in both texture-capturi
ng ability and expressiveness.
****
Learning Robust Representations with Information Bottleneck and Memory Network f
or RGB-D-based Gesture Recognition
Yunan Li, Huizhou Chen, Guanwen Feng, Qiguang Miao; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 20968-20978
Although previous RGB-D-based gesture recognition methods have shown promising p
erformance, researchers often overlook the interference of task-irrelevant cues
like illumination and background. These unnecessary factors are learned together
 with the predictive ones by the network and hinder accurate recognition. In thi
s paper, we propose a convenient and analytical framework to learn a robust feat

ure representation that is impervious to gesture-irrelevant factors. Based on the Information Bottleneck theory, two rules of Sufficiency and Compactness are derived to develop a new information-theoretic loss function, which cultivates a more sufficient and compact representation from the feature encoding and mitigates the impact of gesture-irrelevant information. To highlight the predictive information, we further integrate a memory network. Using our proposed content-based and contextual memory addressing scheme, we weaken the nuisances while preserving the task-relevant information, providing guidance for refining the feature representation. Experiments conducted on three public datasets demonstrate that our approach leads to a better feature representation and achieves better performance than state-of-the-art methods.
****
P1AC: Revisiting Absolute Pose From a Single Affine Correspondence
Jonathan Ventura, Zuzana Kukelova, Torsten Sattler, Dániel Baráth; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19751-19761
Affine correspondences have traditionally been used to improve feature matching over wide baselines. While recent work has successfully used affine correspondences to solve various relative camera pose estimation problems, less attention has been given to their use in absolute pose estimation. We introduce the first general solution to the problem of estimating the pose of a calibrated camera given a single observation of an oriented point and an affine correspondence. The advantage of our approach (P1AC) is that it requires only a single correspondence, in comparison to the traditional point-based approach (P3P), significantly reducing the combinatorics in robust estimation. P1AC provides a general solution that removes restrictive assumptions made in prior work and is applicable to large-scale image-based localization. We propose a minimal solution to the P1AC problem and evaluate our novel solver on synthetic data, showing its numerical stability and performance under various types of noise. On standard image-based localization benchmarks we show that P1AC achieves more accurate results than the widely used P3P algorithm. Code for our method is available at https://github.com/jonathanventura/P1AC/.
****
LAN-HDR: Luminance-based Alignment Network for High Dynamic Range Video Reconstruction
Haesoo Chung, Nam Ik Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12760-12769
As demands for high-quality videos continue to rise, high-resolution and high-dynamic range (HDR) imaging techniques are drawing attention. To generate an HDR video from low dynamic range (LDR) images, one of the critical steps is the motion compensation between LDR frames, for which most existing works employed the optical flow algorithm. However, these methods suffer from flow estimation errors when saturation or complicated motions exist. In this paper, we propose an end-to-end HDR video composition framework, which aligns LDR frames in the feature space and then merges aligned features into an HDR frame, without relying on pixel-domain optical flow. Specifically, we propose a luminance-based alignment network for HDR (LAN-HDR) consisting of an alignment module and a hallucination module. The alignment module aligns a frame to the adjacent reference by evaluating luminance-based attention, excluding color information. The hallucination module generates sharp details, especially for washed-out areas due to saturation. The aligned and hallucinated features are then blended adaptively to complement each other. Finally, we merge the features to generate a final HDR frame. In training, we adopt a temporal loss, in addition to frame reconstruction losses, to enhance temporal consistency and thus reduce flickering. Extensive experiments demonstrate that our method performs better or comparable to state-of-the-art methods on several benchmarks. Codes are available at https://github.com/haesoochung/LAN-HDR.
****
Dancing in the Dark: A Benchmark towards General Low-light Video Enhancement
Huiyuan Fu, Wenkai Zheng, Xicong Wang, Jiaxuan Wang, Heng Zhang, Huadong Ma; Pro

ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
3, pp. 12877-12886
Low-light video enhancement is a challenging task with broad applications. Howev
er, current research in this area is limited by the lack of high-quality benchma
rk datasets. To address this issue, we design a camera system and collect a high
-quality low-light video dataset with multiple exposures and cameras. Our datase
t provides dynamic video pairs with pronounced camera motion and strict spatial
alignment. To achieve general low-light video enhancement, we also propose a nov
el Retinex-based method named Light Adjustable Network (LAN). LAN iteratively re
fines the illumination and adaptively adjusts it under varying lighting conditio
ns, leading to visually appealing results even in diverse real-world scenarios.
The extensive experiments demonstrate the superiority of our low-light video dat
aset and enhancement method. Our dataset and code will be publicly available.
****
RED-PSM: Regularization by Denoising of Partially Separable Models for Dynamic I
maging
Berk Iskender, Marc L. Klasky, Yoram Bresler; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2023, pp. 10595-10604
Dynamic imaging involves the recovery of a time-varying 2D or 3D object at each
time instant using its undersampled measurements. In particular, in dynamic tomo
graphy, only a single projection at a single view angle may be available at a ti
me, making the problem severely ill-posed. In this work, we propose an approach,
RED-PSM, which combines for the first time two powerful techniques to address t
his challenging imaging problem. The first, are partially separable models, whic
h have been used to introduce a low-rank prior for the spatio-temporal object. T
he second is the recent Regularization by Denoising (RED), which provides a flex
ible framework to exploit the impressive performance of state-of-the-art image d
enoising algorithms, for various inverse problems. We propose a partially separa
ble objective with RED and an optimization scheme with variable splitting and AD
MM. Our objective is proved to converge to a value corresponding to a stationary
point satisfying the first-order optimality conditions. Convergence is accelera
ted by a particular projection-domain-based initialization. We demonstrate the p
erformance and computational improvements of our proposed RED-PSM with a learned
image denoiser by comparing it to a recent deep-prior-based method TD-DIP. Alth
ough the emphasis is on dynamic tomography, we also demonstrate the performance
advantages of RED-PSM in a dynamic cardiac MRI setting.
****
Unsupervised Manifold Linearizing and Clustering
Tianjiao Ding, Shengbang Tong, Kwan Ho Ryan Chan, Xili Dai, Yi Ma, Benjamin D. H
aeffele; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2023, pp. 5450-5461
We consider the problem of simultaneously clustering and learning a linear repre
sentation of data lying close to a union of low-dimensional manifolds, a fundame
ntal task in machine learning and computer vision. When the manifolds are assume
d to be linear subspaces, this reduces to the classical problem of subspace clus
tering, which has been studied extensively over the past two decades. Unfortunat
ely, many real-world datasets such as natural images can not be well approximate
d by linear subspaces. On the other hand, numerous works have attempted to learn
an appropriate transformation of the data, such that data is mapped from a unio
n of general non-linear manifolds to a union of linear subspaces (with points fr
om the same manifold being mapped to the same subspace). However, many existing
works have limitations such as assuming knowledge of the membership of samples t
o clusters, requiring high sampling density, or being shown theoretically to lea
rn trivial representations. In this paper, we propose to optimize the Maximal Co
ding Rate Reduction metric with respect to both the data representation and a no
vel doubly stochastic cluster membership, inspired by state-of-the-art subspace
clustering results. We give a parameterization of such a representation and memb
ership, allowing efficient mini-batching and one-shot initialization. Experiment
s on CIFAR-10, -20, -100, and TinyImageNet-200 datasets show that the proposed m
ethod is much more accurate and scalable than state-of-the-art deep clustering m

ethods, and further learns a latent linear representation of the data.
****
Lossy and Lossless (L2) Post-training Model Size Compression
Yumeng Shi, Shihao Bai, Xiuying Wei, Ruihao Gong, Jianlei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17546-17556
Deep neural networks have delivered remarkable performance and have been widely used in various visual tasks. However, their huge sizes cause significant inconvenience for transmission and storage. Many previous studies have explored model size compression. However, these studies often approach various lossy and lossless compression methods in isolation, leading to challenges in achieving high compression ratios efficiently. This work proposes a post-training model size compression method that combines lossy and lossless compression in a unified way. We first propose a unified parametric weight transformation, which ensures different lossy compression methods can be performed jointly in a post-training manner. Then, a dedicated differentiable counter is introduced to guide the optimization of lossy compression to arrive at a more suitable point for later lossless compression. Additionally, our method can easily control a desired global compression ratio and allocate adaptive ratios for different layers. Finally, our method can achieve a stable 10 times compression ratio without sacrificing accuracy and a 20 times compression ratio with minor accuracy loss in a short time. Our code is available at https://github.com/ModelTC/L2_Compression.
****
C2ST: Cross-Modal Contextualized Sequence Transduction for Continuous Sign Language Recognition
Huaiwen Zhang, Zihang Guo, Yang Yang, Xin Liu, De Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21053-21062
Continuous Sign Language Recognition (CSLR) aims to transcribe the signs of an untrimmed video into written words or glosses. The mainstream framework for CSLR consists of a spatial module for visual representation learning, a temporal module aggregating the local and global temporal information of frame sequence, and the connectionist temporal classification (CTC) loss, which aligns video features with gloss sequence. Unfortunately, the language prior implicit in the gloss sequence is ignored throughout the modeling process. Furthermore, the contextualization of glosses is further ignored in alignment learning, as CTC makes an independence assumption between glosses. In this paper, we propose a Cross-modal Contextualized Sequence Transduction (C2ST) for CSLR, which effectively incorporates the knowledge of gloss sequence into the process of video representation learning and sequence transduction. Specifically, we introduce a cross-modal context learning framework for CSLR, in which the linguistic features of gloss sequences is extracted by a language model, and recurrently integrate with visual features for video modelling. Moreover, we introduce the contextualized sequence transduction loss that incorporates the contextual information of gloss sequences in label prediction, without making any independence assumptions between the glosses. Our method sets the new state of the art on three widely used large-scale sign language recognition datasets: Phoenix-2014, Phoenix-2014-T, and CSL-Daily. On CSL-Daily, our approach achieves an absolute gain of 4.9% WER compared to the best published results.
****
ObjectFusion: Multi-modal 3D Object Detection with Object-Centric Fusion
Qi Cai, Yingwei Pan, Ting Yao, Chong-Wah Ngo, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18067-18076
Recent progress on multi-modal 3D object detection has featured BEV (Bird-Eye-View) based fusion, which effectively unifies both LiDAR point clouds and camera images in a shared BEV space. Nevertheless, it is not trivial to perform camera-to-BEV transformation due to the inherently ambiguous depth estimation of each pixel, resulting in spatial misalignment between these two multi-modal features. Moreover, such transformation also inevitably leads to projection distortion of camera image features in BEV space. In this paper, we propose a novel Object-centric Fusion (ObjectFusion) paradigm, which completely gets rid of camera-to-BEV t

ransformation during fusion to align object-centric features across different modalities for 3D object detection. ObjectFusion first learns three kinds of modality-specific feature maps (i.e., voxel, BEV, and image features) from LiDAR point clouds and its BEV projections, camera images. Then a set of 3D object proposals are produced from the BEV features via a heatmap-based proposal generator. Next, the 3D object proposals are reprojected back to voxel, BEV, and image spaces. We leverage voxel and RoI pooling to generate spatially aligned object-centric features for each modality. All the object-centric features of three modalities are further fused at object level, which is finally fed into the detection heads. Extensive experiments on nuScenes dataset demonstrate the superiority of our ObjectFusion, by achieving 69.8% mAP on nuScenes validation set and improving BEVFusion by 1.3%.
****
D-IF: Uncertainty-aware Human Digitization via Implicit Distribution Field
Xueting Yang, Yihao Luo, Yuliang Xiu, Wei Wang, Hao Xu, Zhaoxin Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9122-9132
Realistic virtual humans play a crucial role in numerous industries, such as metaverse, intelligent healthcare, and self-driving simulation. But creating them on a large scale with high levels of realism remains a challenge. The utilization of deep implicit function sparks a new era of image-based 3D clothed human reconstruction, enabling pixel-aligned shape recovery with fine details. Subsequently, the vast majority of works locate the surface by regressing the deterministic implicit value for each point. However, should all points be treated equally regardless of their proximity to the surface? In this paper, we propose replacing the implicit value with an adaptive uncertainty distribution, to differentiate between points based on their distance to the surface. This simple "value to distribution" transition yields significant improvements on nearly all the baselines. Furthermore, qualitative results demonstrate that the models trained using our uncertainty distribution loss, can capture more intricate wrinkles, and realistic limbs. Code and models are available for research purposes at https://github.com/psyai-net/D-IF_release.
****
MMVP: Motion-Matrix-Based Video Prediction
Yiqi Zhong, Luming Liang, Ilya Zharkov, Ulrich Neumann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4273-4283
A central challenge of video prediction lies where the system has to reason the object's future motion from image frames while simultaneously maintaining the consistency of its appearance across frames. This work introduces an end-to-end trainable two-stream video prediction framework, Motion-Matrix-based Video Prediction (MMVP), to tackle this challenge. Unlike previous methods that usually handle motion prediction and appearance maintenance within the same set of modules, MMVP decouples motion and appearance information by constructing appearance-agnostic motion matrices. The motion matrices represent the temporal similarity of each and every pair of feature patches in the input frames, and are the sole input of the motion prediction module in MMVP. This design improves video prediction in both accuracy and efficiency, and reduces the model size. Results of extensive experiments demonstrate that MMVP outperforms state-of-the-art systems on public data sets by non-negligible large margins (approx. 1 db in PSNR, UCF Sports) in significantly smaller model sizes (84% the size or smaller). Please refer to https://github.com/Kay1794/MMVP-motion-matrix-based-video-prediction for the official code and the datasets used in this paper.
****
Human Preference Score: Better Aligning Text-to-Image Models with Human Preference
Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2096-2105
Recent years have witnessed a rapid growth of deep generative models, with text-to-image models gaining significant attention from the public. However, existing models often generate images that do not align well with human preferences, suc

h as awkward combinations of limbs and facial expressions. To address this issue, we collect a dataset of human choices on generated images from the Stable Foundation Discord channel. Our experiments demonstrate that current evaluation metrics for generative models do not correlate well with human choices. Thus, we train a human preference classifier with the collected dataset and derive a Human Preference Score (HPS) based on the classifier. Using HPS, we propose a simple yet effective method to adapt Stable Diffusion to better align with human preferences. Our experiments show that HPS outperforms CLIP in predicting human choices and has good generalization capability toward images generated from other models. By tuning Stable Diffusion with the guidance of HPS, the adapted model is able to generate images that are more preferred by human users. The project page is available here: https://tgxs002.github.io/align_sd_web/.
****
Guided Motion Diffusion for Controllable Human Motion Synthesis
Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, Siyu Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2151-2162
Denoising diffusion models have shown great promise in human motion synthesis conditioned on natural language descriptions. However, integrating spatial constraints, such as pre-defined motion trajectories and obstacles, remains a challenge despite being essential for bridging the gap between isolated human motion and its surrounding environment. To address this issue, we propose Guided Motion Diffusion (GMD), a method that incorporates spatial constraints into the motion generation process. Specifically, we propose an effective feature projection scheme that manipulates motion representation to enhance the coherency between spatial information and local poses. Together with a new imputation formulation, the generated motion can reliably conform to spatial constraints such as global motion trajectories. Furthermore, given sparse spatial constraints (e.g. sparse keyframes), we introduce a new dense guidance approach to turn a sparse signal, which is susceptible to being ignored during the reverse steps, into denser signals to guide the generated motion to the given constraints. Our extensive experiments justify the development of \methodname, which achieves a significant improvement over state-of-the-art methods in text-based motion generation while allowing control of the synthesized motions with spatial constraints.
****
AffordPose: A Large-Scale Dataset of Hand-Object Interactions with Affordance-Driven Hand Pose
Juntao Jian, Xiuping Liu, Manyi Li, Ruizhen Hu, Jian Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14713-14724
How human interact with objects depends on the functional roles of the target objects, which introduces the problem of affordance-aware hand-object interaction. It requires a large number of human demonstrations for the learning and understanding of plausible and appropriate hand-object interactions. In this work, we present AffordPose, a large-scale dataset of hand-object interactions with affordance-driven hand pose. We first annotate the specific part-level affordance labels for each object, e.g. twist, pull, handle-grasp, etc, instead of the general intents such as use or handover, to indicate the purpose and guide the localization of the hand-object interactions. The fine-grained hand-object interactions reveal the influence of hand-centered affordances on the detailed arrangement of the hand poses, yet also exhibit a certain degree of diversity. We collect a total of 26.7K hand-object interactions, each including the 3D object shape, the part-level affordance label, and the manually adjusted hand poses. The comprehensive data analysis shows the common characteristics and diversity of hand-object interactions per affordance via the parameter statistics and contacting computation. We also conduct experiments on the tasks of hand-object affordance understanding and affordance-oriented hand-object interaction generation, to validate the effectiveness of our dataset in learning the fine-grained hand-object interactions. Project page: https://github.com/GentlesJan/AffordPose .
****
Locomotion-Action-Manipulation: Synthesizing Human-Scene Interactions in Complex

3D Environments
Jiye Lee, Hanbyul Joo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9663-9674
Synthesizing interaction-involved human motions has been challenging due to the high complexity of 3D environments and the diversity of possible human behaviors within. We present LAMA, Locomotion-Action-MAnipulation, to synthesize natural and plausible long term human movements in complex indoor environments. The key motivation of LAMA is to build a unified framework to encompass a series of everyday motions including locomotion, scene interaction, and object manipulation. Unlike existing methods that require motion data "paired" with scanned 3D scenes for supervision, we formulate the problem as a test-time optimization by using human motion capture data only for synthesis. LAMA leverages a reinforcement learning framework coupled with motion matching algorithm for optimization, and further exploits a motion editing framework via manifold learning to cover possible variations in interaction and manipulation. Throughout extensive experiments, we demonstrate that LAMA outperforms previous approaches in synthesizing realistic motions in various challenging scenarios.
****
NDDepth: Normal-Distance Assisted Monocular Depth Estimation
Shuwei Shao, Zhongcai Pei, Weihai Chen, Xingming Wu, Zhengguo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7931-7940
Monocular depth estimation has drawn widespread attention from the vision community due to its broad applications. In this paper, we propose a novel physics (geometry)-driven deep learning framework for monocular depth estimation by assuming that 3D scenes are constituted by piece-wise planes. Particularly, we introduce a new normal-distance head that outputs pixel-level surface normal and plane-to-origin distance for deriving depth at each position. Meanwhile, the normal and distance are regularized by a developed plane-aware consistency constraint. We further integrate an additional depth head to improve the robustness of the proposed framework. To fully exploit the strengths of these two heads, we develop an effective contrastive iterative refinement module that refines depth in a complementary manner according to the depth uncertainty. Extensive experiments indicate that the proposed method exceeds previous state-of-the-art competitors on the NYU-Depth-v2, KITTI and SUN RGB-D datasets. Notably, it ranks 1st among all submissions on the KITTI depth prediction online benchmark at the submission time. The source code is available at https://github.com/ShuweiShao/NDDepth.
****
Sequential Texts Driven Cohesive Motions Synthesis with Natural Transitions
Shuai Li, Sisi Zhuang, Wenfeng Song, Xinyu Zhang, Hejia Chen, Aimin Hao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9498-9508
The intelligent synthesis/generation of daily-life motion sequences is fundamental and urgently needed for many VR/metaverse-related applications. However, existing approaches commonly focus on monotonic motion generation (e.g., walking, jumping, etc.) based on single instruction-like text, which is still not intelligent enough and can't meet practical demands. To this end, we propose a cohesive human motion sequence synthesis framework based on free-form sequential texts while ensuring semantic connection and natural transitions between adjacent motions. At the technical level, we explore the local-to-global semantic features of previous and current texts to extract relevant information. This information is used to guide the framework in understanding the semantics of the current moment. Moreover, we propose learnable tokens to adaptively learn the influence range of the previous motions towards natural transitions. These tokens can be trained to encode the relevant information into well-designed transition loss. To demonstrate the efficacy of our method, we conduct extensive experiments and comprehensive evaluations on the public dataset as well as a new dataset produced by us. All the experiments confirm that our method outperforms the state-of-the-art methods in terms of semantic matching, realism, and transition fluency. Our project is public available. https://druthrie.github.io/sequential-texts-to-motion/

****
Efficient Converted Spiking Neural Network for 3D and 2D Classification
Yuxiang Lan, Yachao Zhang, Xu Ma, Yanyun Qu, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9211-9220
Spiking Neural Networks (SNNs) have attracted enormous research interest due to their low-power and biologically plausible nature. Existing ANN-SNN conversion methods can achieve lossless conversion by converting a well-trained Artificial Neural Network (ANN) into an SNN. However, converted SNN requires a large amount of time steps to achieve competitive performance with the well-trained ANN, which means a large latency. In this paper, we propose an efficient unified ANN-SNN conversion method for point cloud classification and image classification to significantly reduce the time step to meet the fast and lossless ANN-SNN transformation. Specifically, we first adaptively adjust the threshold according to the activation state of spiking neurons, ensuring a certain proportion of spiking neurons are activated at each time step to reduce the time for accumulation of membrane potential. Next, we use an adaptive firing mechanism to enlarge the range of spiking output, getting more discrimination features in short time steps. Extensive experimental results on challenging point cloud and image datasets demonstrate that the suggested approach significantly outmatches state-of-the-art ANN-SNN conversion based methods.
****
Eulerian Single-Photon Vision
Shantanu Gupta, Mohit Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10465-10476
Single-photon sensors measure light signals at the finest possible resolution -- individual photons. These sensors introduce two major challenges in the form of strong Poisson noise and extremely large data acquisition rates, which are also inherited by downstream computer vision tasks. Previous work has largely focused on solving the image reconstruction problem first and then using off-the-shelf methods for downstream tasks, but the most general solutions that account for motion are costly and not scalable to large data volumes produced by single-photon sensors. This work forgoes the image reconstruction problem. Instead, we demonstrate computationally light-weight phase-based algorithms for the tasks of edge detection and motion estimation. These methods directly process the raw single-photon data as a 3D volume with a bank of velocity-tuned filters, achieving speed-ups of more than two orders of magnitude compared to explicit reconstruction-based methods. Project webpage: https://wisionlab.com/project/eulerian-single-photon-vision/
****
Adaptive Calibrator Ensemble: Navigating Test Set Difficulty in Out-of-Distribution Scenarios
Yuli Zou, Weijian Deng, Liang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19333-19342
Model calibration usually requires optimizing some parameters (e.g., temperature) w.r.t an objective function like negative log-likelihood. This work uncovers a significant aspect often overlooked that the objective function is influenced by calibration set difficulty: the ratio of misclassified to correctly classified samples. If a test set has a drastically different difficulty level from the calibration set, a phenomenon out-of-distribution (OOD) data often exhibit: the optimal calibration parameters of the two datasets would be different, rendering an optimal calibrator on the calibration set suboptimal on the OOD test set and thus degraded calibration performance. With this knowledge, we propose a simple and effective method named adaptive calibrator ensemble (ACE) to calibrate OOD datasets whose difficulty is usually higher than the calibration set. Specifically, two calibration functions are trained, one for in-distribution data (low difficulty), and the other for severely OOD data (high difficulty). To achieve desirable calibration on a new OOD dataset, ACE uses an adaptive weighting method that strikes a balance between the two extreme functions. When plugged in, ACE generally improves the performance of a few state-of-the-art calibration schemes on a series of OOD benchmarks. Importantly, such improvement does not come at the co

st of the in-distribution calibration performance. Project Website: https://gith
ub.com/insysgroup/Adaptive-Calibrators-Ensemble.git.
****
Contrastive Learning Relies More on Spatial Inductive Bias Than Supervised Learn
ing: An Empirical Study
Yuanyi Zhong, Haoran Tang, Jun-Kun Chen, Yu-Xiong Wang; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2023, pp. 16327-16336
Though self-supervised contrastive learning (CL) has shown its potential to achi
eve state-of-the-art accuracy without any supervision, its behavior still remain
s under investigated by academia. Different from most previous work that underst
ands CL from learning objectives, we focus on an unexplored yet natural aspect:
the spatial inductive bias which seems to be implicitly exploited via data augme
ntations in CL. We design an experiment to study the reliance of CL on such spat
ial inductive bias, by destroying the global or local spatial structures of imag
e with global or local patch shuffling, and comparing the performance drop betwe
en experiments on original and corrupted dataset to quantify the reliance of cer
tain inductive bias. We also use the uniformity of feature space to further rese
arch on how CL-pre-trained model behave with the corrupted dataset. Our results
and analysis show that CL has a much higher reliance on spatial inductive bias t
han SL, regardless of specific CL algorithm or backbones, opening a new directio
n for studying the behavior of CL.
****
DiffuMask: Synthesizing Images with Pixel-level Annotations for Semantic Segment
ation Using Diffusion Models
Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, Chunhua Shen; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 120
6-1217
Collecting and annotating images with pixel-wise labels is time-consuming and la
borious. In contrast, synthetic data can be freely available using a generative
model (e.g., DALL-E, Stable Diffusion). In this paper, we show that it is possib
le to automatically obtain accurate semantic masks of synthetic images generated
by the pre-trained Stable Diffusion, which uses only text-image pairs during tr
aining. Our approach, called DiffuMask, exploits the potential of the cross-atte
ntion map between text and image, which is natural and seamless to extend the te
xt-driven
image synthesis to semantic mask generation. DiffuMask uses text-guided cross-a
ttention information to localize class/word-specific regions, which are combined
with practical techniques to create a novel high-resolution and class-discrimin
ative pixel-wise mask. The methods help to reduce data collection and annotation
costs obviously. Experiments demonstrate that the existing segmentation methods
trained on synthetic data of DiffuMask can achieve a
competitive performance over the counterpart of real data (VOC 2012, Cityscapes
). For some classes (e.g., bird), DiffuMask presents promising performance, clos
e to the state-of-the-art result of real data (within 3% mIoU gap). Moreover, in
the open-vocabulary segmentation (zero-shot) setting,
DiffuMask achieves a new SOTA result on Unseen class of VOC 2012.
****
NSF: Neural Surface Fields for Human Modeling from Monocular Depth
Yuxuan Xue, Bharat Lal Bhatnagar, Riccardo Marin, Nikolaos Sarafianos, Yuanlu Xu
, Gerard Pons-Moll, Tony Tung; Proceedings of the IEEE/CVF International Confere
nce on Computer Vision (ICCV), 2023, pp. 15049-15060
Obtaining personalized 3D animatable avatars from a monocular camera has several
real world applications in gaming, virtual try-on, animation, and VR/XR, etc. H
owever, it is very challenging to model dynamic and fine-grained clothing deform
ations from such sparse data. Existing methods for modeling 3D humans from depth
data have limitations in terms of computational efficiency, mesh coherency, and
flexibility in resolution and topology. For instance, reconstructing shapes usi
ng implicit functions and extracting explicit meshes per frame is computationall
y expensive and cannot ensure coherent meshes across frames. Moreover, predictin
g per-vertex deformations on a pre-designed human template with a discrete surfa

ce lacks flexibility in resolution and topology. To overcome these limitations, we propose a novel method 'NSF: Neural Surface Fields' for modeling 3D clothed humans from monocular depth. NSF defines a neural field solely on the base surface which models a continuous and flexible displacement field. NSF can be adapted to the base surface with different resolution and topology without retraining at inference time. Compared to existing approaches, our method eliminates the expensive per-frame surface extraction while maintaining mesh coherency, and is capable of reconstructing meshes with arbitrary resolution without retraining. To foster research in this direction, we release our code in project page at: https://yuxuan-xue.com/nsf.
****
Unaligned 2D to 3D Translation with Conditional Vector-Quantized Code Diffusion using Transformers
Abril Corona-Figueroa, Sam Bond-Taylor, Neelanjan Bhowmik, Yona Falinie A. Gaus, Toby P. Breckon, Hubert P. H. Shum, Chris G. Willcocks; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14585-14594
Generating 3D images of complex objects conditionally from a few 2D views is a difficult synthesis problem, compounded by issues such as domain gap and geometric misalignment. For instance, a unified framework such as Generative Adversarial Networks cannot achieve this unless they explicitly define both a domain-invariant and geometric-invariant joint latent distribution, whereas Neural Radiance Fields are generally unable to handle both issues as they optimize at the pixel level. By contrast, we propose a simple and novel 2D to 3D synthesis approach based on conditional diffusion with vector-quantized codes. Operating in an information-rich code space enables high-resolution 3D synthesis via full-coverage attention across the views. Specifically, we generate the 3D codes (e.g. for CT images) conditional on previously generated 3D codes and the entire codebook of two 2D views (e.g. 2D X-rays). Qualitative and quantitative results demonstrate state-of-the-art performance over specialized methods across varied evaluation criteria, including fidelity metrics such as density, coverage, and distortion metrics for two complex volumetric imagery datasets from in real-world scenarios.
****
DMNet: Delaunay Meshing Network for 3D Shape Representation
Chen Zhang, Ganzhangqin Yuan, Wenbing Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14418-14428
Recently, there has been a growing interest in learning-based explicit methods due to their ability to respect the original input and preserve details. However, the connectivity on complex structures is still difficult to infer due to the limited local shape perception, resulting in artifacts and non-watertight triangles. In this paper, we present a novel learning-based method with Delaunay triangulation to achieve high-precision reconstruction. We model the Delaunay triangulation as a dual graph, extract local geometric information from the points, and embed it into the structural representation of Delaunay triangulation in an organic way, benefiting fine-grained details reconstruction. To encourage neighborhood information interaction of edges and nodes in the graph, we introduce a local graph iteration algorithm, which is a variant of graph neural network. Moreover, a geometric constraint loss further improves the classification of tetrahedrons. Benefiting from our fully local network, a scaling strategy is designed to enable large-scale reconstruction. Experiments show that our method yields watertight and high-quality meshes. Especially for some thin structures and sharp edges, our method shows better performance than the current state-of-the-art methods. Furthermore, it has a strong adaptability to point clouds of different densities.
****
StyleDomain: Efficient and Lightweight Parameterizations of StyleGAN for One-shot and Few-shot Domain Adaptation
Aibek Alanov, Vadim Titov, Maksim Nakhodnov, Dmitry Vetrov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2184-2194
Domain adaptation of GANs is a problem of fine-tuning GAN models pretrained on a large dataset (e.g. StyleGAN) to a specific domain with few samples (e.g. paint

ing faces, sketches, etc.). While there are many methods that tackle this problem in different ways, there are still many important questions that remain unanswered. In this paper, we provide a systematic and in-depth analysis of the domain adaptation problem of GANs, focusing on the StyleGAN model. We perform a detailed exploration of the most important parts of StyleGAN that are responsible for adapting the generator to a new domain depending on the similarity between the source and target domains. As a result of this study, we propose new efficient and lightweight parameterizations of StyleGAN for domain adaptation. Particularly, we show that there exist directions in StyleSpace (StyleDomain directions) that are sufficient for adapting to similar domains. For dissimilar domains, we propose Affine+ and AffineLight+ parameterizations that allows us to outperform existing baselines in few-shot adaptation while having significantly less training parameters. Finally, we examine StyleDomain directions and discover their many surprising properties that we apply for domain mixing and cross-domain image morphing. Source code can be found at https://github.com/AIRI-Institute/StyleDomain.
****
RankMixup: Ranking-Based Mixup Training for Network Calibration
Jongyoun Noh, Hyekang Park, Junghyup Lee, Bumsub Ham; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1358-1368
Network calibration aims to accurately estimate the level of confidences, which is particularly important for employing deep neural networks in real-world systems. Recent approaches leverage mixup to calibrate the network's predictions during training. However, they do not consider the problem that mixtures of labels in mixup may not accurately represent the actual distribution of augmented samples. In this paper, we present RankMixup, a novel mixup-based framework alleviating the problem of the mixture of labels for network calibration. To this end, we propose to use an ordinal ranking relationship between raw and mixup-augmented samples as an alternative supervisory signal to the label mixtures for network calibration. We hypothesize that the network should estimate a higher level of confidence for the raw samples than the augmented ones (Fig.1). To implement this idea, we introduce a mixup-based ranking loss (MRL) that encourages lower confidences for augmented samples compared to raw ones, maintaining the ranking relationship. We also propose to leverage the ranking relationship among multiple mixup-augmented samples to further improve the calibration capability. Augmented samples with larger mixing coefficients are expected to have higher confidences and vice versa (Fig.1). That is, the order of confidences should be aligned with that of mixing coefficients. To this end, we introduce a novel loss, M-NDCG, in order to reduce the number of misaligned pairs of the coefficients and confidences. Extensive experimental results on standard benchmarks for network calibration demonstrate the effectiveness of RankMixup.
****
Body Knowledge and Uncertainty Modeling for Monocular 3D Human Body Reconstruction
Yufei Zhang, Hanjing Wang, Jeffrey O. Kephart, Qiang Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9020-9032
While 3D body reconstruction methods have made remarkable progress recently, it remains difficult to acquire the sufficiently accurate and numerous 3D supervisions required for training. In this paper, we propose KNOWN, a framework that effectively utilizes body KNOWledge and uNcertainty modeling to compensate for insufficient 3D supervisions. KNOWN exploits a comprehensive set of generic body constraints derived from well-established body knowledge. These generic constraints precisely and explicitly characterize the reconstruction plausibility and enable 3D reconstruction models to be trained without any 3D data. Moreover, existing methods typically use images from multiple datasets during training, which can result in data noise (e.g., inconsistent joint annotation) and data imbalance (e.g., minority images representing unusual poses or captured from challenging camera views). KNOWN solves these problems through a novel probabilistic framework that models both aleatoric and epistemic uncertainty. Aleatoric uncertainty is encoded in a robust Negative Log-Likelihood (NLL) training loss, while epistemic uncertainty is used to guide model refinement. Experiments demonstrate that KNOW

N's body reconstruction outperforms prior weakly-supervised approaches, particularly on the challenging minority images.
****
Randomized Quantization: A Generic Augmentation for Data Agnostic Self-supervised Learning
Huimin Wu, Chenyang Lei, Xiao Sun, Peng-Shuai Wang, Qifeng Chen, Kwang-Ting Cheng, Stephen Lin, Zhirong Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16305-16316
Self-supervised representation learning follows a paradigm of withholding some part of the data and tasking the network to predict it from the remaining part. Among many techniques, data augmentation lies at the core for creating the information gap. Towards this end, masking has emerged as a generic and powerful tool where content is withheld along the sequential dimension, e.g., spatial in images, temporal in audio, and syntactic in language. In this paper, we explore the orthogonal channel dimension for generic data augmentation by exploiting precision redundancy. The data for each channel is quantized through a non-uniform quantizer, with the quantized value sampled randomly within randomly sampled quantization bins. From another perspective, quantization is analogous to channel-wise masking, as it removes the information within each bin, but preserves the information across bins. Our approach significantly surpasses existing generic data augmentation methods, while showing on par performance against modality- specific augmentations. We comprehensively evaluate our approach on vision, audio, 3D point clouds, as well as the DABS benchmark which is comprised of various data modalities. The code is available at https: //github.com/microsoft/random_quantize.
****
Learning to Generate Semantic Layouts for Higher Text-Image Correspondence in Text-to-Image Synthesis
Minho Park, Jooyeol Yun, Seunghwan Choi, Jaegul Choo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7591-7600
Existing text-to-image generation approaches have set high standards for photorealism and text-image correspondence, largely benefiting from web-scale text-image datasets, which can include up to 5 billion pairs. However, text-to-image generation models trained on domain-specific datasets, such as urban scenes, medical images, and faces, still suffer from low text-image correspondence due to the lack of text-image pairs. Additionally, collecting billions of text-image pairs for a specific domain can be time-consuming and costly. Thus, ensuring high text-image correspondence without relying on web-scale text-image datasets remains a challenging task. In this paper, we present a novel approach for enhancing text-image correspondence by leveraging available semantic layouts. Specifically, we propose a Gaussian-categorical diffusion process that simultaneously generates both images and corresponding layout pairs. Our experiments reveal that we can guide text-to-image generation models to be aware of the semantics of different image regions, by training the model to generate semantic labels for each pixel. We demonstrate that our approach achieves higher text-image correspondence compared to existing text-to-image generation approaches in the Multi-Modal CelebA-HQ and the Cityscapes dataset, where text-image pairs are scarce.
****
Neural Radiance Field with LiDAR maps
MingFang Chang, Akash Sharma, Michael Kaess, Simon Lucey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17914-17923
We address outdoor Neural Radiance Fields (NeRF) with LiDAR maps. Existing NeRF methods usually require specially collected hypersampled source views and do not perform well with the open source camera-LiDAR datasets - significantly limiting the approach's practical utility. In this paper, we demonstrate an approach that allows for these datasets to be utilized for high quality neural renderings. Our design leverages 1) LiDAR sensors for strong 3D geometry priors that significantly improve the ray sampling locality, and 2) Conditional Adversarial Networks (cGANs) to recover image details since aggregating embeddings from imperfect LiDAR maps causes artifacts in the synthesized images. Our experiments show that while NeRF baselines produce either noisy or blurry results on Argoverse 2, the

images synthesized by our system not only outperform baselines in image quality metrics under both clean and noisy conditions, but also obtain closer Detectron2 results to the ground truth images. Furthermore, to show the substantial applic ability of our system, we demonstrate that our system can be used in data augmen tation for training a pose regression network and multi-season view synthesis. O ur dataset and code will be released
****
AREA: Adaptive Reweighting via Effective Area for Long-Tailed Classification
Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, Weiping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19277-19287
Large-scale data from real-world usually follow a long-tailed distribution (i.e. , a few majority classes occupy plentiful training data, while most minority cla sses have few samples), making the hyperplanes heavily skewed to the minority cl asses. Traditionally, reweighting is adopted to make the hyperplanes fairly spli t the feature space, where the weights are designed according to the number of s amples.
However, we find that the number of samples in a class can not accurately measu re the size of its spanned space, especially for the majority class, where the s ize of its spanned space is usually larger than the samples' number because of t he high diversity. Therefore, weights designed based on the samples' number will still compress the space of minority classes. In this paper, we reconsider rewe ighting from a totally new perspective of analyzing the spanned space of each cl ass. We argue that, besides statistical numbers, relations between samples are a lso significant for sufficiently depicting the spanned space. Consequently, we e stimate the size of the spanned space for each category, namely effective area, by detailedly analyzing its samples' distribution. By treating samples of a clas s as identically distributed random variables and analyzing their correlations, a simple and non-parametric formula is derived to estimate the effective area. T hen, the weight simply calculated inversely proportional to the effective area o f each class is adopted to achieve fairer training. Note that our weights are mo re flexible as they can be adaptively adjusted along with the optimizing feature s during training.
Experiments on four long-tailed datasets show that the proposed weights outperf orm the state-of-the-art reweighting methods. Moreover, our method can also achi eve better results on statistically balanced CIFAR-10/100. Code is available at https://github.com/xiaohua-chen/AREA.
****
Erasing Concepts from Diffusion Models
Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, David Bau; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2426-2436
Motivated by concerns that large-scale diffusion models can produce undesirable output such as sexually explicit content or copyrighted artistic styles, we stud y erasure of specific concepts from diffusion model weights. We propose a fine-t uning method that can erase a visual concept from a pre-trained diffusion model, given only the name of the style and using negative guidance as a teacher. We b enchmark our method against previous approaches that remove sexually explicit co ntent and demonstrate its effectiveness, performing on par with Safe Latent Diff usion and censored training. To evaluate artistic style removal, we conduct expe riments erasing five modern artists from the network and conduct a user study to assess the human perception of the removed styles. Unlike previous methods, our approach can remove concepts from a diffusion model permanently rather than mod ifying the output at the inference time, so it cannot be circumvented even if a user has access to model weights. Our code, data, and results are available at e rasing.baulab.info
****
Fully Attentional Networks with Self-emerging Token Labeling
Bingyin Zhao, Zhiding Yu, Shiyi Lan, Yutao Cheng, Anima Anandkumar, Yingjie Lao, Jose M. Alvarez; Proceedings of the IEEE/CVF International Conference on Comput

er Vision (ICCV), 2023, pp. 5585-5595
Recent studies indicate that Vision Transformers (ViTs) are robust against out-of-distribution scenarios. In particular, the Fully Attentional Network (FAN) - a family of ViT backbones, has achieved state-of-the-art robustness. In this paper, we revisit the FAN models and improve their pre-training with a self-emerging token labeling (STL) framework. Our method contains a two-stage training framework. Specifically, we first train a FAN token labeler (FAN-TL) to generate semantically meaningful patch token labels, followed by a FAN student model training stage that uses both the token labels and the original class label. With the proposed STL framework, our best model based on FAN-L-Hybrid (77.3M parameters) achieves 84.8% Top-1 accuracy and 42.1% mCE on ImageNet-1K and ImageNet-C, and sets a new state-of-the-art for ImageNet-A (46.1%) and ImageNet-R (56.6%) without using extra data, outperforming the original FAN counterpart by significant margins. The proposed framework also demonstrates significantly enhanced performance on downstream tasks such as semantic segmentation, with up to 1.7% improvement in robustness over the counterpart model.
****
ACTIVE: Towards Highly Transferable 3D Physical Camouflage for Universal and Robust Vehicle Evasion
Naufal Suryanto, Yongsu Kim, Harashta Tatimma Larasati, Hyoeun Kang, Thi-Thu-Huong Le, Yoonyoung Hong, Hunmin Yang, Se-Yoon Oh, Howon Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4305-4314
Adversarial camouflage has garnered attention for its ability to attack object detectors from any viewpoint by covering the entire object's surface. However, universality and robustness in existing methods often fall short as the transferability aspect is often overlooked, thus restricting their application only to a specific target with limited performance. To address these challenges, we present Adversarial Camouflage for Transferable and Intensive Vehicle Evasion (ACTIVE), a state-of-the-art physical camouflage attack framework designed to generate universal and robust adversarial camouflage capable of concealing any 3D vehicle from detectors. Our framework incorporates innovative techniques to enhance universality and robustness, including a refined texture rendering that enables common texture application to different vehicles without being constrained to a specific texture map, a novel stealth loss that renders the vehicle undetectable, and a smooth and camouflage loss to enhance the naturalness of the adversarial camouflage. Our extensive experiments on 15 different models show that ACTIVE consistently outperforms existing works on various public detectors, including the latest YOLOv7. Notably, our universality evaluations reveal promising transferability to other vehicle classes, tasks (segmentation models), and the real world, not just other vehicles.
****
Learning Adaptive Neighborhoods for Graph Neural Networks
Avishkar Saha, Oscar Mendez, Chris Russell, Richard Bowden; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22541-22550
Graph convolutional networks (GCNs) enable end-to-end learning on graph structured data. However, many works assume a given graph structure. When the input graph is noisy or unavailable, one approach is to construct or learn a latent graph structure. These methods typically fix the choice of node degree for the entire graph, which is suboptimal. Instead, we propose a novel end-to-end differentiable graph generator which builds graph topologies where each node selects both its neighborhood and its size. Our module can be readily integrated into existing pipelines involving graph convolution operations, replacing the predetermined or existing adjacency matrix with one that is learned, and optimized, as part of the general objective. As such it is applicable to any GCN. We integrate our module into trajectory prediction, point cloud classification and node classification pipelines resulting in improved accuracy over other structure-learning methods across a wide range of datasets and GCN backbones.
****
Equivariant Similarity for Vision-Language Foundation Models

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, Lijuan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11998-12008

This study explores the concept of equivariance in vision-language foundation models (VLMs), focusing specifically on the multimodal similarity function that is not only the major training objective but also the core delivery to support downstream tasks. Unlike the existing image-text similarity objective which only categorizes matched pairs as similar and unmatched as dissimilar, equivariance also requires similarity to vary faithfully according to the semantic changes. This allows VLMs to generalize better to nuanced and unseen multimodal compositions. However, modeling equivariance is challenging as the ground truth of semantic change is difficult to collect. For example, given an image-text pair about a dog, it is unclear to what extent the similarity changes when the pixel is changed from dog to cat? To this end, we propose EqSim, a regularization loss that can be efficiently calculated from any two matched training pairs and easily pluggable into existing image-text retrieval fine-tuning. Meanwhile, to further diagnose the equivariance of VLMs, we present a new challenging benchmark EqBen. Compared to the existing evaluation sets, EqBen is the first to focus on "visual-minimal change". Extensive experiments show the lack of equivariance in current VLMs and validate the effectiveness of EqSim.

****

ReST: A Reconfigurable Spatial-Temporal Graph Model for Multi-Camera Multi-Object Tracking

Cheng-Che Cheng, Min-Xuan Qiu, Chen-Kuo Chiang, Shang-Hong Lai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10051-10060

Multi-Camera Multi-Object Tracking (MC-MOT) utilizes information from multiple views to better handle problems with occlusion and crowded scenes. Recently, the use of graph-based approaches to solve tracking problems has become very popular. However, many current graph-based methods do not effectively utilize information regarding spatial and temporal consistency. Instead, they rely on single-camera trackers as input, which are prone to fragmentation and ID switch errors. In this paper, we propose a novel reconfigurable graph model that first associates all detected objects across cameras spatially before reconfiguring it into a temporal graph for Temporal Association. This two-stage association approach enables us to extract robust spatial and temporal-aware features and address the problem with fragmented tracklets. Furthermore, our model is designed for online tracking, making it suitable for real-world applications. Experimental results show that the proposed graph model is able to extract more discriminating features for object tracking, and our model achieves state-of-the-art performance on several public datasets. Code is available at https://github.com/chengche6230/ReST.

****

Too Large; Data Reduction for Vision-Language Pre-Training

Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3147-3157

This paper examines the problems of severe image-text misalignment and high redundancy in the widely-used large-scale Vision-Language Pre-Training (VLP) datasets. To address these issues, we propose an efficient and straightforward Vision-Language learning algorithm called TL;DR which aims to compress the existing large VLP data into a small, high-quality set. Our approach consists of two major steps. First, a codebook-based encoder-decoder captioner is developed to select representative samples. Second, a new caption is generated to complement the original captions for selected samples, mitigating the text-image misalignment problem while maintaining uniqueness. As the result, TL;DR enables us to reduce the large dataset into a small set of high-quality data, which can serve as an alternative pre-training dataset. This algorithm significantly speeds up the time-consuming pretraining process. Specifically, TL;DR can compress the mainstream VLP datasets at a high ratio, e.g., reduce well-cleaned CC3M dataset from 2.8M to 0.67M (24%) and noisy YFCC15M from 15M to 2.5M (16.7%). Extensive experiments with

 three popular VLP models over seven downstream tasks show that VLP model traine
d on the compressed dataset provided by TL;DR can perform similar or even better
 results compared with training on the full-scale dataset.
****
Make-It-3D: High-fidelity 3D Creation from A Single Image with Diffusion Prior
Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, Dong Chen;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2023, pp. 22819-22829
In this work, we investigate the problem of creating high-fidelity 3D content fr
om only a single image. This is inherently challenging: it essentially involves
estimating the underlying 3D geometry while hallucinating unseen textures. To ad
dress this challenge, we leverage prior knowledge in a well-trained 2D diffusion
 model to serve as a 3D-aware supervision for 3D creation. Our proposed method,
Make-It-3D, employs a two-stage optimization pipeline: the first stage optimizes
 a neural radiance field with constraints from the reference image and diffusion
 prior; the second stage builds textured point clouds from the coarse model and
further enhances the textures with diffusion prior leveraging the availability o
f high-quality textures from the reference image. Extensive experiments show tha
t our method achieves a clear improvement over previous works, displaying faithf
ul reconstruction and impressive visual quality. Our method presents the first a
ttempt to achieve high-quality 3D creation from a single image for general objec
ts, and enables various applications such as text-to-3D creation and texture edi
ting.
****
Towards Deeply Unified Depth-aware Panoptic Segmentation with Bi-directional Gui
dance Learning
Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Bin Luo, Jun-Yan He, Jin-Peng Lan
, Yifeng Geng, Xuansong Xie; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2023, pp. 4111-4121
Depth-aware panoptic segmentation is an emerging topic in computer vision which
combines semantic and geometric understanding for more robust scene interpretati
on. Recent works pursue unified frameworks to tackle this challenge but mostly s
till treat it as two individual learning tasks, which limits their potential for
 exploring cross-domain information. We propose a deeply unified framework for d
epth-aware panoptic segmentation, which performs joint segmentation and depth es
timation both in a per-segment manner with identical object queries. To narrow t
he gap between the two tasks, we further design a geometric query enhancement me
thod, which is able to integrate scene geometry into object queries using latent
 representations. In addition, we propose a bi-directional guidance learning app
roach to facilitate cross-task feature learning by taking advantage of their mut
ual relations. Our method sets the new state of the art for depth-aware panoptic
 segmentation on both Cityscapes-DVPS and SemKITTI-DVPS datasets. Moreover, our
guidance learning approach is shown to deliver performance improvement even unde
r incomplete supervision labels. Code and models are available at https://github
.com/jwh97nn/DeepDPS.
****
Taxonomy Adaptive Cross-Domain Adaptation in Medical Imaging via Optimization Tr
ajectory Distillation
Jianan Fan, Dongnan Liu, Hang Chang, Heng Huang, Mei Chen, Weidong Cai; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 21174-21184
The success of automated medical image analysis depends on large-scale and exper
t-annotated training sets. Unsupervised domain adaptation (UDA) has been raised
as a promising approach to alleviate the burden of labeled data collection. Howe
ver, they generally operate under the closed-set adaptation setting assuming an
identical label set between the source and target domains, which is over-restric
tive in clinical practice where new classes commonly exist across datasets due t
o taxonomic inconsistency. While several methods have been presented to tackle b
oth domain shifts and incoherent label sets, none of them take into account the
common characteristics of the two issues and consider the learning dynamics alon

g network training. In this work, we propose optimization trajectory distillation, a unified approach to address the two technical challenges from a new perspective. It exploits the low-rank nature of gradient space and devises a dual-stream distillation algorithm to regularize the learning dynamics of insufficiently annotated domain and classes with the external guidance obtained from reliable sources. Our approach resolves the issue of inadequate navigation along network optimization, which is the major obstacle in the taxonomy adaptive cross-domain adaptation scenario. We evaluate the proposed method extensively on several tasks towards various endpoints with clinical significance. The results demonstrate its effectiveness and improvements over previous methods.
****
DiffTAD: Temporal Action Detection with Proposal Denoising Diffusion
Sauradip Nag, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10362-10374
We propose a new formulation of temporal action detection (TAD) with denoising diffusion, DiffTAD in short. Taking as input random temporal proposals, it can yield action proposals accurately given an untrimmed long video. This presents a generative modeling perspective, against previous discriminative learning manners. This capability is achieved by first diffusing the ground-truth proposals to random ones (i.e, the forward/noising process) and then learning to reverse the noising process (i.e, the backward/denoising process). Concretely, we establish the denoising process in the Transformer decoder (e.g, DETR) by introducing a temporal location query design with faster convergence in training. We further propose a cross-step selective conditioning algorithm for inference acceleration. Extensive evaluations on ActivityNet and THUMOS show that our DiffTAD achieves top performance compared to previous art alternatives. The code is available at https://github.com/sauradip/DiffusionTAD.
****
Ray Conditioning: Trading Photo-consistency for Photo-realism in Multi-view Image Generation
Eric Ming Chen, Sidhanth Holalkere, Ruyu Yan, Kai Zhang, Abe Davis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23242-23251
Multi-view image generation attracts particular attention these days due to its promising 3D-related applications, e.g., image viewpoint editing. Most existing methods follow a paradigm where a 3D representation is first synthesized, and then rendered into 2D images to ensure photo-consistency across viewpoints. However, such explicit bias for photo-consistency sacrifices photo-realism, causing geometry artifacts and loss of fine-scale details when these methods are applied to edit real images. To address this issue, we propose ray conditioning, a geometry-free alternative that relaxes the photo-consistency constraint. Our method generates multi-view images by conditioning a 2D GAN on a light field prior. With explicit viewpoint control, state-of-the-art photo-realism and identity consistency, our method is particularly suited for the viewpoint editing task.
****
SCOB: Universal Text Understanding via Character-wise Supervised Contrastive Learning with Online Text Rendering for Bridging Domain Gap
Daehee Kim, Yoonsik Kim, DongHyun Kim, Yumin Lim, Geewook Kim, Taeho Kil; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19562-19573
Inspired by the great success of language model (LM)-based pre-training, recent studies in visual document understanding have explored LM-based pre-training methods for modeling text within document images. Among them, pre-training that reads all text from an image has shown promise, but often exhibits instability and even fails when applied to broader domains, such as those involving both visual documents and scene text images. This is a substantial limitation for real-world scenarios, where the processing of text image inputs in diverse domains is essential. In this paper, we investigate effective pre-training tasks in the broader domains and also propose a novel pre-training method called SCOB that leverages

 character-wise supervised contrastive learning with online text rendering to effectively pre-train document and scene text domains by bridging the domain gap. Moreover, SCOB enables weakly supervised learning, significantly reducing annotation costs. Extensive benchmarks demonstrate that SCOB generally improves vanilla pre-training methods and achieves comparable performance to state-of-the-art methods. Our findings suggest that SCOB can be served generally and effectively for read-type pre-training methods. The code will be available at https://github.com/naver-ai/scob.
****
Point-Query Quadtree for Crowd Counting, Localization, and More
Chengxin Liu, Hao Lu, Zhiguo Cao, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1676-1685
We show that crowd counting can be viewed as a decomposable point querying process. This formulation enables arbitrary points as input and jointly reasons whether the points are crowd and where they locate. The querying processing, however, raises an underlying problem on the number of necessary querying points. Too few imply underestimation; too many increase computational overhead. To address this dilemma, we introduce a decomposable structure, i.e., the point-query quadtree, and propose a new counting model, termed Point quEry Transformer (PET). PET implements decomposable point querying via data-dependent quadtree splitting, where each querying point could split into four new points when necessary, thus enabling dynamic processing of sparse and dense regions. Such a querying process yields an intuitive, universal modeling of crowd as both the input and output are interpretable and steerable. We demonstrate the applications of PET on a number of crowd-related tasks, including fully-supervised crowd counting and localization, partial annotation learning, and point annotation refinement, and also report state-of-the-art performance. For the first time, we show that a single counting model can address multiple crowd-related tasks across different learning paradigms. Code is available at https://github.com/cxliu0/PET.
****
Heterogeneous Diversity Driven Active Learning for Multi-Object Tracking
Rui Li, Baopeng Zhang, Jun Liu, Wei Liu, Jian Zhao, Zhu Teng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9932-9941
The existing one-stage multi-object tracking (MOT) algorithms have achieved satisfactory performance benefiting from a large amount of labeled data. However, acquiring plenty of laborious annotated frames is not practical in real applications. To reduce the cost of human annotations, we propose Heterogeneous Diversity driven Active Multi-Object Tracking (HD-AMOT), to infer the most informative frames for any MOT tracker by observing the heterogeneous cues of samples. HD-AMOT defines the diversified informative representation by encoding the geometric and semantic information, and formulates the frame inference strategy as a Markov decision process to learn an optimal sampling policy based on the designed informative representation. Specifically, HD-AMOT consists of a diversified informative representation module as well as an informative frame selection network. The former produces the signal characterizing the diversity and distribution of frames, and the latter receives the signal and conducts multi-frame cooperation to enable batch frame sampling. Extensive experiments conducted on the MOT15, MOT17, MOT20, and Dancetrack datasets demonstrate the efficacy and effectiveness of HD-AMOT. Experiments show that under 50% budget our HD-AMOT can achieve similar or even higher performance as fully-supervised learning.
****
Domain Generalization of 3D Semantic Segmentation in Autonomous Driving
Jules Sanchez, Jean-Emmanuel Deschaud, François Goulette; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18077-18087
Using deep learning, 3D autonomous driving semantic segmentation has become a well-studied subject, with methods that can reach very high performance. Nonetheless, because of the limited size of the training datasets, these models cannot see every type of object and scene found in real-world applications. The ability to be reliable in these various unknown environments is called domain generalizat

ion. Despite its importance, domain generalization is relatively unexplored in the case of 3D autonomous driving semantic segmentation. To fill this gap, this paper presents the first benchmark for this application by testing state-of-the-art methods and discussing the difficulty of tackling Laser Imaging Detection and Ranging (LiDAR) domain shifts. We also propose the first method designed to address this domain generalization, which we call 3DLabelProp. This method relies on leveraging the geometry and sequentiality of the LiDAR data to enhance its generalization performances by working on partially accumulated point clouds. It reaches a mean Intersection over Union (mIoU) of 50.4% on SemanticPOSS and of 55.2% on PandaSet solid-state LiDAR while being trained only on SemanticKITTI, making it the state-of-the-art method for generalization (+5% and +33% better, respectively, than the second best method).
****
HaMuCo: Hand Pose Estimation via Multiview Collaborative Self-Supervised Learning
Xiaozheng Zheng, Chao Wen, Zhou Xue, Pengfei Ren, Jingyu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20763-20773
Recent advancements in 3D hand pose estimation have shown promising results, but its effectiveness has primarily relied on the availability of large-scale annotated datasets, the creation of which is a laborious and costly process. To alleviate the label-hungry limitation, we propose a self-supervised learning framework, HaMuCo, that learns a single view hand pose estimator from multi-view pseudo 2D labels. However, one of the main challenges of self-supervised learning is the presence of noisy labels and the "groupthink" effect from multiple views. To overcome these issues, we introduce a cross-view interaction network that distills the single view estimator by utilizing the cross-view correlated features and enforcing multi-view consistency to achieve collaborative learning. Both the single view estimator and the cross-view interaction network are trained jointly in an end-to-end manner. Extensive experiments show that our method can achieve state-of-the-art performance on multi-view self-supervised hand pose estimation. Furthermore, the proposed cross-view interaction network can also be applied to hand pose estimation from multi-view input and outperforms previous methods under same settings.
****
Efficient Model Personalization in Federated Learning via Client-Specific Prompt Generation
Fu-En Yang, Chien-Yi Wang, Yu-Chiang Frank Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19159-19168
Federated learning (FL) emerges as a decentralized learning framework which trains models from multiple distributed clients without sharing their data to preserve privacy. Recently, large-scale pre-trained models (e.g., Vision Transformer) have shown a strong capability of deriving robust representations. However, the data heterogeneity among clients, the limited computation resources, and the communication bandwidth restrict the deployment of large-scale models in FL frameworks. To leverage robust representations from large-scale models while enabling efficient model personalization for heterogeneous clients, we propose a novel personalized FL framework of client-specific Prompt Generation (pFedPG), which learns to deploy a personalized prompt generator at the server for producing client-specific visual prompts that efficiently adapts frozen backbones to local data distributions. Our proposed framework jointly optimizes the stages of personalized prompt adaptation locally and personalized prompt generation globally. The former aims to train visual prompts that adapt foundation models to each client, while the latter observes local optimization directions to generate personalized prompts for all clients. Through extensive experiments on benchmark datasets, we show that our pFedPG is favorable against state-of-the-art personalized FL methods under various types of data heterogeneity, allowing computation and communication efficient model personalization.
****
Dual Aggregation Transformer for Image Super-Resolution

Zheng Chen, Yulun Zhang, Jinjin Gu, Linghe Kong, Xiaokang Yang, Fisher Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12312-12321
Transformer has recently gained considerable popularity in low-level vision tasks, including image super-resolution (SR). These networks utilize self-attention along different dimensions, spatial or channel, and achieve impressive performance. This inspires us to combine the two dimensions in Transformer for a more powerful representation capability. Based on the above idea, we propose a novel Transformer model, Dual Aggregation Transformer (DAT), for image SR. Our DAT aggregates features across spatial and channel dimensions, in the inter-block and intra-block dual manner. Specifically, we alternately apply spatial and channel self-attention in consecutive Transformer blocks. The alternate strategy enables DAT to capture the global context and realize inter-block feature aggregation. Furthermore, we propose the adaptive interaction module (AIM) and the spatial-gate feed-forward network (SGFN) to achieve intra-block feature aggregation. AIM complements two self-attention mechanisms from corresponding dimensions. Meanwhile, SGFN introduces additional non-linear spatial information in the feed-forward network. Extensive experiments show that our DAT surpasses current methods. Code and models are obtainable at https://github.com/zhengchen1999/DAT.
****
Zero-Shot Spatial Layout Conditioning for Text-to-Image Diffusion Models
Guillaume Couairon, Marlène Careil, Matthieu Cord, Stéphane Lathuilière, Jakob Verbeek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2174-2183
Large-scale text-to-image diffusion models have significantly improved the state of the art in generative image modeling and allow for an intuitive and powerful user interface to drive the image generation process. Expressing spatial constraints, e.g. to position specific objects in particular locations, is cumbersome using text; and current text-based image generation models are not able to accurately follow such instructions. In this paper we consider image generation from text associated with segments on the image canvas, which combines an intuitive natural language interface with precise spatial control over the generated content. We propose ZestGuide, a "ZEro-shot" SegmenTation Guidance approach that can be plugged into pre-trained text-to-image diffusion models, and does not require any additional training. It leverages implicit segmentation maps that can be extracted from cross-attention layers, and uses them to align the generation with input masks. Our experimental results combine high image quality with accurate alignment of generated content with input segmentations, and improve over prior work both quantitatively and qualitatively, including methods that require training on images with corresponding segmentations. Compared to Paint with Words, the previous state-of-the art in image generation with zero-shot segmentation conditioning, we improve by 5 to 10 mIoU points on the COCO dataset with similar FID scores.
****
SegGPT: Towards Segmenting Everything in Context
Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, Tiejun Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1130-1140
We present SegGPT, a generalist model for segmenting everything in context. We unify various segmentation tasks into a generalist in-context learning framework that accommodates different kinds of segmentation data by transforming them into the same format of images. The training of SegGPT is formulated as an in-context coloring problem with random color mapping for each data sample. The objective is to accomplish diverse tasks according to the context, rather than relying on specific colors. After training, SegGPT can perform arbitrary segmentation tasks in images or videos via in-context inference, such as object instance, stuff, part, contour, and text. SegGPT is evaluated on a broad range of tasks, including few-shot semantic segmentation, video object segmentation, semantic segmentation, and panoptic segmentation. Our results show strong capabilities in segmenting in-domain and out-of-domain targets, either qualitatively or quantitatively.

****
Semantify: Simplifying the Control of 3D Morphable Models Using CLIP
Omer Gralnik, Guy Gafni, Ariel Shamir; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14554-14564
We present Semantify: a self-supervised method that utilizes the semantic power of CLIP language-vision foundation model to simplify the control of 3D morphable models. Given a parametric model, training data is created by randomly sampling the model's parameters, creating various shapes and rendering them. The similarity between the output images and a set of word descriptors is calculated in CLIP's latent space. Our key idea is first to choose a small set of semantically meaningful and disentangled descriptors that characterize the 3DMM, and then learn a non-linear mapping from scores across this set to the parametric coefficients of the given 3DMM. The non-linear mapping is defined by training a neural network without a human-in-the-loop. We present results on numerous 3DMMs: body shape models, face shape and expression models, as well as animal shapes. We demonstrate how our method defines a simple slider interface for intuitive modeling, and show how the mapping can be used to instantly fit a 3D parametric body shape to in-the-wild images.
****
From Sky to the Ground: A Large-scale Benchmark and Simple Baseline Towards Real Rain Removal
Yun Guo, Xueyao Xiao, Yi Chang, Shumin Deng, Luxin Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12097-12107
Learning-based image deraining methods have made great progress. However, the lack of large-scale high-quality paired training samples is the main bottleneck to hamper the real image deraining (RID). To address this dilemma and advance RID, we construct a Large-scale High-quality Paired real rain benchmark (LHP-Rain), including 3000 video sequences with 1 million high-resolution (1920*1080) frame pairs. The advantages of the proposed dataset over the existing ones are three-fold: rain with higher-diversity and larger-scale, image with higher-resolution and higher-quality ground-truth. Specifically, the real rains in LHP-Rain not only contain the classical rain streak/veiling/occlusion in the sky, but also the splashing on the ground overlooked by deraining community. Moreover, we propose a novel robust low-rank tensor recovery model to generate the GT with better separating the static background from the dynamic rain. In addition, we design a simple transformer-based single image deraining baseline, which simultaneously utilize the self-attention and cross-layer attention within the image and rain layer with discriminative feature representation. Extensive experiments verify the superiority of the proposed dataset and deraining method over state-of-the-art.
****
Knowledge Restore and Transfer for Multi-Label Class-Incremental Learning
Songlin Dong, Haoyu Luo, Yuhang He, Xing Wei, Jie Cheng, Yihong Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18711-18720
Current class-incremental learning research mainly focuses on single-label classification tasks while multi-label class-incremental learning (MLCIL) with more practical application scenarios is rarely studied. Although there have been many anti-forgetting methods to solve the problem of catastrophic forgetting in single-label class-incremental learning, these methods have difficulty in solving the MLCIL problem due to label absence and information dilution problems. To solve these problems, we propose a Knowledge Restore and Transfer (KRT) framework including a dynamic pseudo-label (DPL) module to solve the label absence problem by restoring the knowledge of old classes to the new data and an incremental cross-attention (ICA) module with session-specific knowledge retention tokens storing knowledge and a unified knowledge transfer token transferring knowledge to solve the information dilution problem. Comprehensive experimental results on MS-COCO and PASCAL VOC datasets demonstrate the effectiveness of our method for improving recognition performance and mitigating forgetting on multi-label class-incremental learning tasks.
****

DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders
Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, Xuansong Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 328-338
Image colorization is a challenging problem due to multi-modal uncertainty and high ill-posedness. Directly training a deep neural network usually leads to incorrect semantic colors and low color richness. While transformer-based methods can deliver better results, they often rely on manually designed priors, suffer from poor generalization ability, and introduce color bleeding effects. To address these issues, we propose DDColor, an end-to-end method with dual decoders for image colorization. Our approach includes a pixel decoder and a query-based color decoder. The former restores the spatial resolution of the image, while the latter utilizes rich visual features to refine color queries, thus avoiding hand-crafted priors. Our two decoders work together to establish correlations between color and multi-scale semantic representations via cross-attention, significantly alleviating the color bleeding effect. Additionally, a simple yet effective colorfulness loss is introduced to enhance the color richness. Extensive experiments demonstrate that DDColor achieves superior performance to existing state-of-the-art works both quantitatively and qualitatively. The codes and models are publicly available.
****
Visual Explanations via Iterated Integrated Attributions
Oren Barkan, ■Yehonatan Elisha■■, Yuval Asher, Amit Eshel, Noam Koenigstein; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2073-2084
We introduce Iterated Integrated Attributions (IIA) - a generic method for explaining the predictions of vision models. IIA employs iterative integration across the input image, the internal representations generated by the model, and their gradients, yielding precise and focused explanation maps. We demonstrate the effectiveness of IIA through comprehensive evaluations across various tasks, datasets, and network architectures. Our results showcase that IIA produces accurate explanation maps, outperforming other state-of-the-art explanation techniques.
****
PanFlowNet: A Flow-Based Deep Network for Pan-Sharpening
Gang Yang, Xiangyong Cao, Wenzhe Xiao, Man Zhou, Aiping Liu, Xun Chen, Deyu Meng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16857-16867
Pan-sharpening aims to generate a high-resolution multispectral (HRMS) image by integrating the spectral information of a low-resolution multispectral (LRMS) image with the texture details of a high-resolution panchromatic (PAN) image. It essentially inherits the ill-posed nature of the super-resolution (SR) task that diverse HRMS images can degrade into an LRMS image. However, existing deep learning-based methods recover only one HRMS image from the LRMS image and PAN image using a deterministic mapping, thus ignoring the diversity of the HRMS image. In this paper, to alleviate this ill-posed issue, we propose a flow-based pan-sharpening network (PanFlowNet) to directly learn the conditional distribution of HRMS image given LRMS image and PAN image instead of learning a deterministic mapping. Specifically, we first transform this unknown conditional distribution into a given Gaussian distribution by an invertible network, and the conditional distribution can thus be explicitly defined. Then, we design an invertible Conditional Affine Coupling Block (CACB) and further build the architecture of PanFlowNet by stacking a series of CACBs. Finally, the PanFlowNet is trained by maximizing the log-likelihood of the conditional distribution given a training set and can then be used to predict diverse HRMS images. The experimental results verify that the proposed PanFlowNet can generate various HRMS images given an LRMS image and a PAN image. Additionally, the experimental results on different kinds of satellite datasets also demonstrate the superiority of our PanFlowNet compared with other state-of-the-art methods both visually and quantitatively. Code is available at Github.
****

Domain Generalization via Balancing Training Difficulty and Model Capability
Xueying Jiang, Jiaxing Huang, Sheng Jin, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18993-19003
Domain generalization (DG) aims to learn domaingeneralizable models from one or multiple source domains that can perform well in unseen target domains. Despite its recent progress, most existing work suffers from the misalignment between the difficulty level of training samples and the capability of contemporarily trained models, leading to over-fitting or under-fitting in the trained generalization model. We design MoDify, a Momentum Difficulty framework that tackles the misalignment by balancing the seesaw between the model's capability and the samples' difficulties along the training process. MoDify consists of two novel designs that collaborate to fight against the misalignment while learning domain-generalizable models. The first is MoDify-based Data Augmentation which exploits an RGB Shuffle technique to generate difficulty-aware training samples on the fly. The second is MoDify-based Network Optimization which dynamically schedules the training samples for balanced and smooth learning with appropriate difficulty. Without bells and whistles, a simple implementation of MoDify achieves superior performance across multiple benchmarks. In addition, MoDify can complement existing methods as a plug-in, and it is generic and can work for different visual recognition tasks.
****
Pairwise Similarity Learning is SimPLE
Yandong Wen, Weiyang Liu, Yao Feng, Bhiksha Raj, Rita Singh, Adrian Weller, Michael J. Black, Bernhard Schölkopf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5308-5318
In this paper, we focus on a general yet important learning problem, pairwise similarity learning (PSL). PSL subsumes a wide range of important applications, such as open-set face recognition, speaker verification, image retrieval and person re-identification. The goal of PSL is to learn a pairwise similarity function assigning a higher similarity score to positive pairs (i.e., a pair of samples with the same label) than to negative pairs (i.e., a pair of samples with different label). We start by identifying a key desideratum for PSL, and then discuss how existing methods can achieve this desideratum. We then propose a surprisingly simple proxy-free method, called SimPLE, which requires neither feature/proxy normalization nor angular margin and yet is able to generalize well in open-set recognition. We apply the proposed method to three challenging PSL tasks: open-set face recognition, image retrieval and speaker verification. Comprehensive experimental results on large-scale benchmarks show that our method performs significantly better than current state-of-the-art methods.
****
GO-SLAM: Global Optimization for Consistent 3D Instant Reconstruction
Youmin Zhang, Fabio Tosi, Stefano Mattoccia, Matteo Poggi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3727-3737
Neural implicit representations have recently demonstrated compelling results on dense Simultaneous Localization And Mapping (SLAM) but suffer from the accumulation of errors in camera tracking and distortion in the reconstruction. Purposely, we present GO-SLAM, a deep-learning-based dense visual SLAM framework globally optimizing poses and 3D reconstruction in real-time. Robust pose estimation is at its core, supported by efficient loop closing and online full bundle adjustment, which optimize per frame by utilizing the learned global geometry of the complete history of input frames. Simultaneously, we update the implicit and continuous surface representation on-the-fly to ensure global consistency of 3D reconstruction. Results on various synthetic and real-world datasets demonstrate that GO-SLAM outperforms state-of-the-art approaches at tracking robustness and reconstruction accuracy. Furthermore, GO-SLAM is versatile and can run with monocular, stereo, and RGB-D input.
****
JOTR: 3D Joint Contrastive Learning with Transformers for Occluded Human Mesh Recovery
Jiahao Li, Zongxin Yang, Xiaohan Wang, Jianxin Ma, Chang Zhou, Yi Yang; Proceedi

ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9110-9121
In this study, we focus on the problem of 3D human mesh recovery from a single image under obscured conditions. Most state-of-the-art methods aim to improve 2D alignment technologies, such as spatial averaging and 2D joint sampling. However, they tend to neglect the crucial aspect of 3D alignment by improving 3D representations. Furthermore, recent methods struggle to separate target human from occlusion or background in crowded scenes as they optimize the 3D space of target human with 3D joint coordinates as local supervision. To address these issues, a desirable method would involve a framework for fusing 2D and 3D features and a strategy for optimizing the 3D space globally. Therefore, this paper presents 3D JOint contrastive learning with TRansformers (JOTR) framework for handling occluded 3D human mesh recovery. Our method includes an encoder-decoder transformer architecture to fuse 2D and 3D representations for achieving 2D&3D aligned results in a coarse-to-fine manner and a novel 3D joint contrastive learning approach for adding explicitly global supervision for the 3D feature space. The contrastive learning approach includes two contrastive losses: joint-to-joint contrast for enhancing the similarity of semantically similar voxels (i.e., human joints), and joint-to-non-joint contrast for ensuring discrimination from others (e.g., occlusions and background). Qualitative and quantitative analyses demonstrate that our method outperforms state-of-the-art competitors on both occlusion-specific and standard benchmarks, significantly improving the reconstruction of occluded humans.
****
CLIP-Driven Universal Model for Organ Segmentation and Tumor Detection
Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, Zongwei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21152-21164
An increasing number of public datasets have shown a marked impact on automated organ segmentation and tumor detection. However, due to the small size and partially labeled problem of each dataset, as well as a limited investigation of diverse types of tumors, the resulting models are often limited to segmenting specific organs/tumors and ignore the semantics of anatomical structures, nor can they be extended to novel domains.
To address these issues, we propose the CLIP-Driven Universal Model, which incorporates text embedding learned from Contrastive Language-Image Pre-training (CLIP) to segmentation models. This CLIP-based label encoding captures anatomical relationships, enabling the model to learn a structured feature embedding and segment 25 organs and 6 types of tumors. The proposed model is developed from an assembly of 14 datasets, using a total of 3,410 CT scans for training and then evaluated on 6,162 external CT scans from 3 additional datasets. We rank first on the Medical Segmentation Decathlon (MSD) public leaderboard and achieve state-of-the-art results on Beyond The Cranial Vault (BTCV). Additionally, the Universal Model is computationally more efficient (6xfaster) compared with dataset-specific models, generalized better to CT scans from varying sites, and shows stronger transfer learning performance on novel tasks.
****
NIR-assisted Video Enhancement via Unpaired 24-hour Data
Muyao Niu, Zhihang Zhong, Yinqiang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10778-10788
Low-light video enhancement in the visible (VIS) range is important yet technically challenging, and it is likely to become more tractable by introducing near-infrared (NIR) information for assistance, which in turn arouses a new challenge on how to obtain appropriate multispectral data for model training. In this paper, we defend the feasibility and superiority of NIR-assisted low-light video enhancement results by using unpaired 24-hour data for the first time, which significantly eases data collection and improves generalization performance on in-the-wild data. By accounting for different physical characteristics between unpaired daytime and nighttime videos, we first propose to turn daytime NIR & VIS into "nighttime mode". Specifically, we design a heuristic yet physics-inspired religh

ting algorithm to produce realistic pseudo nighttime NIR, and use a resampling strategy followed by a noiseGAN for nighttime VIS conversion. We further devise a temporal-aware network for video enhancement that extracts and fuses bi-directional temporal streams and is trained using real daytime videos and pseudo nighttime videos. We capture multi-spectral data using a co-axial camera and contribute Fulltime Multi-Spectral Video Dataset (FMSVD), the first dataset including aligned 24-hour NIR & VIS videos. Compared to alternative methods, we achieve significantly improved video quality as well as generalization ability on in-the-wild data in terms of both evaluation metrics and visual judgment. Codes and Data Available: https://github.com/MyNiuuu/NVEU.
****
FACTS: First Amplify Correlations and Then Slice to Discover Bias
Sriram Yenamandra, Pratik Ramesh, Viraj Prabhu, Judy Hoffman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4794-4804
Computer vision datasets frequently contain spurious correlations between task-relevant labels and (easy to learn) latent task-irrelevant attributes (e.g. context). Models trained on such datasets learn "shortcuts" and underperform on bias-conflicting slices of data where the correlation does not hold. In this work, we study the problem of identifying such slices to inform downstream bias mitigation strategies. We propose First Amplify Correlations and Then Slice (FACTS), wherein we first amplify correlations to fit a simple bias-aligned hypothesis via strongly regularized empirical risk minimization. Next, we perform correlation-aware slicing via mixture modeling in bias-aligned feature space to discover underperforming data slices that capture distinct correlations. Despite its simplicity, our method considerably improves over prior work (by as much as 35% precision@10) in correlation bias identification across a range of diverse evaluation settings. Our code is available at https://github.com/yvsriram/FACTS.
****
Anchor Structure Regularization Induced Multi-view Subspace Clustering via Enhanced Tensor Rank Minimization
Jintian Ji, Songhe Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19343-19352
The tensor-based multi-view subspace clustering algorithms have received widespread attention due to the powerful ability to capture high-order correlation across views. Although such algorithms have achieved remarkable success, they still suffer from three main issues: 1) The extremely high computational complexity makes tensor-based methods difficult to handle large-scale data sets. 2) The commonly used Tensor Nuclear Norm (TNN) treats different singular values equally and under-penalizes the noise components, resulting in a sub-optimal representation tensor. 3) The subspace-based methods usually ignore the local geometric structure of the original data. Being aware of these, we propose Anchor Structure Regularitation Induced Multi-view Subspace Clustering via Enhanced Tensor Rank Minimization (ASR-ETR). Specifically, an anchor representation tensor is constructed by using the anchor representation strategy rather than the self-representation strategy to reduce the time complexity, and an Anchor Structure Regularization (ASR) is employed to enhance the local geometric structure in the learned anchor-representation tensor. We further define an Enhanced Tensor Rank (ETR), which is a tighter surrogate of the tensor rank and more effective to drive the noise out. Moreover, an efficient iterative optimization algorithm is designed to solve the ASR-ETR, which enjoys both linear complexity and favorable convergence. Extensive experimental results on various data sets demonstrate the superiority of the proposed algorithm as compared to state-of-the-art methods.
****
VeRi3D: Generative Vertex-based Radiance Fields for 3D Controllable Human Image Synthesis
Xinya Chen, Jiaxin Huang, Yanrui Bin, Lu Yu, Yiyi Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8986-8997
Unsupervised learning of 3D-aware generative adversarial networks has lately made much progress. Some recent work demonstrates promising results of learning hum

an generative models using neural articulated radiance fields, yet their general ization ability and controllability lag behind parametric human models, i.e., th ey do not perform well when generalizing to novel pose/shape and are not part co ntrollable. To solve these problems, we propose VeRi3D, a generative human verte x-based radiance field parameterized by vertices of the parametric human templat e, SMPL. We map each 3D point to the local coordinate system defined on its neig hboring vertices, and use the corresponding vertex feature and local coordinates for mapping it to color and density values. We demonstrate that our simple appr oach allows for generating photorealistic human images with free control over ca mera pose, human pose, shape, as well as enabling part-level editing.
****
MOSE: A New Dataset for Video Object Segmentation in Complex Scenes
Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, Philip H.S. Torr, Song Bai; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 023, pp. 20224-20234
Video object segmentation (VOS) aims at segmenting a particular object throughou t the entire video clip sequence. The state-of-the-art VOS methods have achieved excellent performance (e.g., 90+% J&F) on existing datasets. However, since the target objects in these existing datasets are usually relatively salient, domin ant, and isolated, VOS under complex scenes has rarely been studied. To revisit VOS and make it more applicable in the real world, we collect a new VOS dataset called coMplex video Object SEgmentation (MOSE) to study the tracking and segmen ting objects in complex environments. MOSE contains 2,149 video clips and 5,200 objects from 36 categories, with 431,725 high-quality object segmentation masks. The most notable feature of MOSE dataset is complex scenes with crowded and occ luded objects. The target objects in the videos are commonly occluded by others and disappear in some frames. To analyze the proposed MOSE dataset, we benchmark 18 existing VOS methods under 4 different settings on the proposed MOSE dataset and conduct comprehensive comparisons. The experiments show that current VOS al gorithms cannot well perceive objects in complex scenes. For example, under the semi-supervised VOS setting, the highest J&F by existing state-of-the-art VOS me thods is only 59.4% on MOSE, much lower than their 90% J&F performance on DAVIS . The results reveal that although excellent performance has been achieved on ex isting benchmarks, there are unresolved challenges under complex scenes and more efforts are desired to explore these challenges in the future.
****
BoMD: Bag of Multi-label Descriptors for Noisy Chest X-ray Classification
Yuanhong Chen, Fengbei Liu, Hu Wang, Chong Wang, Yuyuan Liu, Yu Tian, Gustavo Ca rneiro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21284-21295
Deep learning methods have shown outstanding classification accuracy in medical imaging problems, which is largely attributed to the availability of large-scale datasets manually annotated with clean labels. However, given the high cost of such manual annotation, new medical imaging classification problems may need to rely on machine-generated noisy labels extracted from radiology reports. Indeed, many Chest X-Ray (CXR) classifiers have been modelled from datasets with noisy labels, but their training procedure is in general not robust to noisy-label sam ples, leading to sub-optimal models. Furthermore, CXR datasets are mostly multi- label, so current multi-class noisy-label learning methods cannot be easily adap ted. In this paper, we propose a new method designed for noisy multi-label CXR l earning, which detects and smoothly re-labels noisy samples from the dataset to be used in the training of common multi-label classifiers. The proposed method o ptimises a bag of multi-label descriptors (BoMD) to promote their similarity wit h the semantic descriptors produced by language models from multi-label image an notations. Our experiments on noisy multi-label training sets and clean testing sets show that our model has state-of-the-art accuracy and robustness in many CX R multi-label classification benchmarks, including a new benchmark that we propo se to systematically assess noisy multi-label methods.
****
Mask-Attention-Free Transformer for 3D Instance Segmentation

Xin Lai, Yuhui Yuan, Ruihang Chu, Yukang Chen, Han Hu, Jiaya Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3693-3703

Recently, transformer-based methods have dominated 3D instance segmentation, where mask attention is commonly involved. Specifically, object queries are guided by the initial instance masks in the first cross-attention, and then iteratively refine themselves in a similar manner. However, we observe that the mask-attention pipeline usually leads to slow convergence due to low-recall initial instance masks. Therefore, we abandon the mask attention design and resort to an auxiliary center regression task instead. Through center regression, we effectively overcome the low-recall issue and perform cross-attention by imposing positional prior. To reach this goal, we develop a series of position-aware designs. First, we learn a spatial distribution of 3D locations as the initial position queries. They spread over the 3D space densely, and thus can easily capture the objects in a scene with a high recall. Moreover, we present relative position encoding for the cross-attention and iterative refinement for more accurate position queries. Experiments show that our approach converges 4x faster than existing work, sets a new state of the art on ScanNetv2 3D instance segmentation benchmark, and also demonstrates superior performance across various datasets. Code and models are available at https://github.com/dvlab-research/Mask-Attention-Free-Transformer.

****

SHIFT3D: Synthesizing Hard Inputs For Tricking 3D Detectors

Hongge Chen, Zhao Chen, Gregory P. Meyer, Dennis Park, Carl Vondrick, Ashish Shrivastava, Yuning Chai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8493-8503

We present SHIFT3D, a differentiable pipeline for generating 3D shapes that are structurally plausible yet challenging to 3D object detectors. In safety-critical applications like autonomous driving, discovering such novel challenging objects can offer insight into unknown vulnerabilities of 3D detectors. By representing objects with a signed distanced function (SDF), we show that gradient error signals allow us to smoothly deform the shape or pose of a 3D object in order to confuse a downstream 3D detector. Importantly, the objects generated by SHIFT3D physically differ from the baseline object yet retain a semantically recognizable shape. Our approach provides interpretable failure modes for modern 3D object detectors, and can aid in preemptive discovery of potential safety risks within 3D perception systems before these risks become critical failures.

****

EgoLoc: Revisiting 3D Object Localization from Egocentric Videos with Visual Queries

Jinjie Mai, Abdullah Hamdi, Silvio Giancola, Chen Zhao, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 45-57

With the recent advances in video and 3D understanding, novel 4D spatio-temporal methods fusing both concepts have emerged. Towards this direction, the Ego4D Episodic Memory Benchmark proposed a task for Visual Queries with 3D Localization (VQ3D). Given an egocentric video clip and an image crop depicting a query object, the goal is to localize the 3D position of the center of that query object with respect to the camera pose of a query frame. Current methods tackle the problem of VQ3D by unprojecting the 2D localization results of the sibling task Visual Queries with 2D Localization (VQ2D) into 3D predictions. Yet, we point out that the low number of camera poses caused by camera re-localization from previous VQ3D methods severally hinders their overall success rate. In this work, we formalize a pipeline (we dub EgoLoc) that better entangles 3D multiview geometry with 2D object retrieval from egocentric videos. Our approach involves estimating more robust camera poses and aggregating multi-view 3D displacements by leveraging the 2D detection confidence, which enhances the success rate of object queries and leads to a significant improvement in the VQ3D baseline performance. Specifically, our approach achieves an overall success rate of up to 87.12%, which sets a new state-of-the-art result in the VQ3D task. We provide a comprehensive emp

irical analysis of the VQ3D task and existing solutions, and highlight the remaining challenges in VQ3D. The code is available at https://github.com/Wayne-Mai/EgoLoc.
****
Coordinate Transformer: Achieving Single-stage Multi-person Mesh Recovery from Videos
Haoyuan Li, Haoye Dong, Hanchao Jia, Dong Huang, Michael C. Kampffmeyer, Liang Lin, Xiaodan Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8744-8753
Multi-person 3D mesh recovery from videos is a critical first step towards automatic perception of group behavior in virtual reality, physical therapy and beyond. However, existing approaches rely on multi-stage paradigms, where the person detection and tracking stages are performed in a multi-person setting, while temporal dynamics are only modeled for one person at a time. Consequently, their performance is severely limited by the lack of inter-person interactions in the spatial-temporal mesh recovery, as well as by detection and tracking defects. To address these challenges, we propose the Coordinate transFormer (CoordFormer) that directly models multi-person spatial-temporal relations and simultaneously performs multi-mesh recovery in an end-to-end manner. Instead of partitioning the feature map into coarse-scale patch-wise tokens, CoordFormer leverages a novel Coordinate-Aware Attention to preserve pixel-level spatial-temporal coordinate information. Additionally, we propose a simple, yet effective Body Center Attention mechanism to fuse position information. Extensive experiments on the 3DPW dataset demonstrate that CoordFormer significantly improves the state-of-the-art, outperforming the previously best results by 4.2%, 8.8% and 4.7% according to the MPJPE, PAMPJPE, and PVE metrics, respectively, while being 40% faster than recent video-based approaches. The released code can be found at https://github.com/Li-Hao-yuan/CoordFormer.
****
FLatten Transformer: Vision Transformer using Focused Linear Attention
Dongchen Han, Xuran Pan, Yizeng Han, Shiji Song, Gao Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5961-5971
The quadratic computation complexity of self-attention has been a persistent challenge when applying Transformer models to vision tasks. Linear attention, on the other hand, offers a much more efficient alternative with its linear complexity by approximating the Softmax operation through carefully designed mapping functions. However, current linear attention approaches either suffer from significant performance degradation or introduce additional computation overhead from the mapping functions. In this paper, we propose a novel Focused Linear Attention module to achieve both high efficiency and expressiveness. Specifically, we first analyze the factors contributing to the performance degradation of linear attention from two perspectives: the focus ability and feature diversity. To overcome these limitations, we introduce a simple yet effective mapping function and an efficient rank restoration module to enhance the expressiveness of self-attention while maintaining low computation complexity. Extensive experiments show that our linear attention module is applicable to a variety of advanced vision Transformers, and achieves consistently improved performances on multiple benchmarks. Code is available at https://github.com/LeapLabTHU/FLatten-Transformer.
****
Q-Diffusion: Quantizing Diffusion Models
Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, Kurt Keutzer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17535-17545
Diffusion models have achieved great success in image synthesis through iterative noise estimation using deep neural networks. However, the slow inference, high memory consumption, and computation intensity of the noise estimation model hinder the efficient adoption of diffusion models. Although post-training quantization (PTQ) is considered a go-to compression method for other tasks, it does not work out-of-the-box on diffusion models. We propose a novel PTQ method specifically tailored towards the unique multi-timestep pipeline and model architecture o

f the diffusion models, which compresses the noise estimation network to acceler
ate the generation process. We identify the key difficulty of diffusion model qu
antization as the changing output distributions of noise estimation networks ove
r multiple time steps and the bimodal activation distribution of the shortcut la
yers within the noise estimation network. We tackle these challenges with timest
ep-aware calibration and split shortcut quantization in this work. Experimental
results show that our proposed method is able to quantize full-precision uncondi
tional diffusion models into 4-bit while maintaining comparable performance (sma
ll FID change of at most 2.34 compared to >100 for traditional PTQ) in a trainin
g-free manner. Our approach can also be applied to text-guided image generation,
 where we can run stable diffusion in 4-bit weights with high generation quality
 for the first time.
****
Robustifying Token Attention for Vision Transformers
Yong Guo, David Stutz, Bernt Schiele; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2023, pp. 17557-17568
Despite the success of vision transformers (ViTs), they still suffer from signif
icant drops in accuracy in the presence of common corruptions, such as noise or
blur. Interestingly, we observe that the attention mechanism of ViTs tends to re
ly on few important tokens, a phenomenon we call token overfocusing. More critic
ally, these tokens are not robust to corruptions, often leading to highly diverg
ing attention patterns. In this paper, we intend to alleviate this overfocusing
issue and make attention more stable through two general techniques: First, our
Token-aware Average Pooling (TAP) module encourages the local neighborhood of ea
ch token to take part in the attention mechanism. Specifically, TAP learns avera
ge pooling schemes for each token such that the information of potentially impor
tant tokens in the neighborhood can adaptively be taken into account. Second, we
 force the output tokens to aggregate information from a diverse set of input to
kens rather than focusing on just a few by using our Attention Diversification L
oss (ADL). We achieve this by penalizing high cosine similarity between the atte
ntion vectors of different tokens. In experiments, we apply our methods to a wid
e range of transformer architectures and improve robustness significantly. For e
xample, we improve corruption robustness on ImageNet-C by 2.4% while improving a
ccuracy by 0.4% based on state-of-the-art robust architecture FAN. Also, when fi
ne-tuning on semantic segmentation tasks, we improve robustness on CityScapes-C
by 2.4% and ACDC by 3.0%. Our code is available at https://github.com/guoyongcs/
TAPADL.
****
Boosting Positive Segments for Weakly-Supervised Audio-Visual Video Parsing
Kranthi Kumar Rachavarapu, Rajagopalan A. N.; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2023, pp. 10192-10202
In this paper, we address the problem of weakly supervised Audio-Visual Video Pa
rsing (AVVP), where the goal is to temporally localize events that are audible o
r visible and simultaneously classify them into known event categories. This is
a challenging task, as we only have access to the video-level event labels durin
g training but need to predict event labels at the segment level during evaluati
on. Existing multiple-instance learning (MIL) based methods use a form of attent
ive pooling over segment-level predictions. These methods only optimize for a su
bset of most discriminative segments that satisfy the weak-supervision constrain
ts, which miss identifying positive segments. To address this, we focus on impro
ving the proportion of positive segments detected in a video. To this end, we mo
del the number of positive segments in a video as a latent variable and show tha
t it can be modeled as Poisson binomial distribution over segment-level predicti
ons, which can be computed exactly. Given the absence of fine-grained supervisio
n, we propose an Expectation-Maximization approach to learn the model parameters
 by maximizing the evidence lower bound (ELBO). We iteratively estimate the mini
mum positive segments in a video and refine them to capture more positive segmen
ts. We conducted extensive experiments on AVVP tasks to evaluate the effectivene
ss of our proposed approach, and the results clearly demonstrate that it increas
es the number of positive segments captured compared to existing methods. Additi

onally, our experiments on Temporal Action Localization (TAL) demonstrate the po
tential of our method for generalization to similar MIL tasks.
****
ADNet: Lane Shape Prediction via Anchor Decomposition
Lingyu Xiao, Xiang Li, Sen Yang, Wankou Yang; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2023, pp. 6404-6413
In this paper, we revisit the limitations of anchor-based lane detection methods
, which have predominantly focused on fixed anchors that stem from the edges of
the image, disregarding their versatility and quality. To overcome the inflexibi
lity of anchors, we decompose them into learning the heat map of starting points
 and their associated directions. This decomposition removes the limitations on
the starting point of anchors, making our algorithm adaptable to different lane
types in various datasets. To enhance the quality of anchors, we introduce the L
arge Kernel Attention (LKA) for Feature Pyramid Network (FPN). This significantl
y increases the receptive field, which is crucial in capturing the sufficient co
ntext as lane lines typically run throughout the entire image. We have named our
 proposed system the Anchor Decomposition Network (ADNet). Additionally, we prop
ose the General Lane IoU (GLIoU) loss, which significantly improves the performa
nce of ADNet in complex scenarios. Experimental results on three widely used lan
e detection benchmarks, VIL-100, CULane, and TuSimple, demonstrate that our appr
oach outperforms the state-of-the-art methods on VIL-100 and exhibits competitiv
e accuracy on CULane and TuSimple. Code and models will be released on https://g
ithub.com/ Sephirex-X/ADNet.
****
UniSeg: A Unified Multi-Modal LiDAR Segmentation Network and the OpenPCSeg Codeb
ase
Youquan Liu, Runnan Chen, Xin Li, Lingdong Kong, Yuchen Yang, Zhaoyang Xia, Yeqi
 Bai, Xinge Zhu, Yuexin Ma, Yikang Li, Yu Qiao, Yuenan Hou; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21662-2167
3
Point-, voxel-, and range-views are three representative forms of point clouds.
All of them have accurate 3D measurements but lack color and texture information
. RGB images are a natural complement to these point cloud views and fully utili
zing the comprehensive information of them benefits more robust perceptions. In
this paper, we present a unified multi-modal LiDAR segmentation network, termed
UniSeg, which leverages the information of RGB images and three views of the poi
nt cloud, and accomplishes semantic segmentation and panoptic segmentation simul
taneously. Specifically, we first design the Learnable cross-Modal Association (
LMA) module to automatically fuse voxel-view and range-view features with image
features, which fully utilize the rich semantic information of images and are ro
bust to calibration errors. Then, the enhanced voxel-view and range-view feature
s are transformed to the point space, where three views of point cloud features
are further fused adaptively by the Learnable cross-View Association module (LVA
). Notably, UniSeg achieves promising results in three public benchmarks, i.e.,
SemanticKITTI, nuScenes, and Waymo Open Dataset (WOD); it ranks 1st on two chall
enges of two benchmarks, including the LiDAR semantic segmentation challenge of
nuScenes and panoptic segmentation challenges of SemanticKITTI. Besides, we cons
truct the OpenPCSeg codebase, which is the largest and most comprehensive outdoo
r LiDAR segmentation codebase. It contains most of the popular outdoor LiDAR seg
mentation algorithms and provides reproducible implementations. The OpenPCSeg co
debase will be made publicly available at https://github.com/PJLab-ADG/PCSeg.
****
Sign Language Translation with Iterative Prototype
Huijie Yao, Wengang Zhou, Hao Feng, Hezhen Hu, Hao Zhou, Houqiang Li; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
15592-15601
This paper presents IP-SLT, a simple yet effective framework for sign language t
ranslation (SLT). Our IP-SLT adopts a recurrent structure and enhances the seman
tic representation (prototype) of the input sign language video via an iterative
 refinement manner. Our idea mimics the behavior of human reading, where a sente

nce can be digested repeatedly, till reaching accurate understanding. Technically, IP-SLT consists of feature extraction, prototype initialization, and iterative prototype refinement. The initialization module generates the initial prototype based on the visual feature extracted by the feature extraction module. Then, the iterative refinement module leverages the cross-attention mechanism to polish the previous prototype by aggregating it with the original video feature. Through repeated refinement, the prototype finally converges to a more stable and accurate state, leading to a fluent and appropriate translation. In addition, to leverage the sequential dependence of prototypes, we further propose an iterative distillation loss to compress the knowledge of the final iteration into previous ones. As the autoregressive decoding process is executed only once in inference, our IP-SLT is ready to improve various SLT systems with acceptable overhead. Extensive experiments are conducted on public benchmarks to demonstrate the effectiveness of the IP-SLT.
****
Pixel-Wise Contrastive Distillation
Junqiang Huang, Zichao Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16359-16369
We present a simple but effective pixel-level self-supervised distillation framework friendly to dense prediction tasks. Our method, called Pixel-Wise Contrastive Distillation (PCD), distills knowledge by attracting the corresponding pixels from student's and teacher's output feature maps. PCD includes a novel design called SpatialAdaptor which "reshapes" a part of the teacher network while preserving the distribution of its output features. Our ablation experiments suggest that this reshaping behavior enables more informative pixel-to-pixel distillation. Moreover, we utilize a plug-in multi-head self-attention module that explicitly relates the pixels of student's feature maps to enhance the effective receptive field, leading to a more competitive student. PCD outperforms previous self-supervised distillation methods on various dense prediction tasks. A backbone of ResNet-18-FPN distilled by PCD achieves 37.4 AP-bbox and 34.0 AP-mask on COCO dataset using the detector of Mask R-CNN. We hope our study will inspire future research on how to pre-train a small model friendly to dense prediction tasks in a self-supervised fashion.
****
Efficient Deep Space Filling Curve
Wanli Chen, Xufeng Yao, Xinyun Zhang, Bei Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17525-17534
Space-filling curves (SFCs) act as a linearization approach to map data in higher dimensional space to lower dimensional space, which is used comprehensively in computer vision, such as image/point cloud compression, hashing and etc. Currently, researchers formulate the problem of searching for an optimal SFC to the problem of finding a single Hamiltonian circuit on the image grid graph. Existing methods adopt graph neural networks (GNN) for SFC search. By modeling the pixel grid as a graph, they first adopt GNN to predict the edge weights and then generate a minimum spanning tree (MST) based on the predictions, which is further used to construct the SFC. However, GNN-based methods suffer from high computational costs and memory footprint usage. Besides, MST generation is un-differentiable, which is infeasible to optimize via gradient descent. To remedy these issues, we propose a GNN-based SFC-search framework with a tailored algorithm that largely reduces computational cost of GNN.

Additionally, we propose a siamese network learning scheme to optimize DNN-based models in an end-to-end fashion.

Extensive experiments show that our proposed method outperforms both DNN-based methods and traditional SFCs, e.g. Hilbert curve, by a large margin on various benchmarks.
****
GlueGen: Plug and Play Multi-modal Encoders for X-to-image Generation
Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, Ran Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23085-23096

Text-to-image (T2I) models based on diffusion processes have achieved remarkable
success in controllable image generation using user-provided captions. However,
the tight coupling between the current text encoder and image decoder in T2I mo
dels makes it challenging to replace or upgrade. Such changes often require mass
ive fine-tuning or even training from scratch with the prohibitive expense. To a
ddress this problem, we propose GlueGen, which applies a newly proposed GlueNet
model to align features from single-modal or multi-modal encoders with the laten
t space of an existing T2I model. The approach introduces a new training objecti
ve that leverages parallel corpora to align the representation spaces of differe
nt encoders. Empirical results show that GlueNet can be trained efficiently and
enables various capabilities beyond previous state-of-the-art models: 1) multili
ngual language models such as XLM-Roberta can be aligned with existing T2I model
s, allowing for the generation of high-quality images from captions beyond Engli
sh; 2) GlueNet can align multi-modal encoders such as AudioCLIP with the Stable
Diffusion model, enabling sound-to-image generation; 3) it can also upgrade the
current text encoder of the latent diffusion model for challenging case generati
on. By the alignment of various feature representations, the GlueNet allows for
flexible and efficient integration of new functionality into existing T2I models
and sheds light on X-to-image (X2I) generation.
****
Humans in 4D: Reconstructing and Tracking Humans with Transformers
Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, Jitendr
a Malik; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2023, pp. 14783-14794
We present an approach to reconstruct humans and track them over time. At the co
re of our approach, we propose a fully "transformerized" version of a network fo
r human mesh recovery. This network, HMR 2.0, advances the state of the art and
shows the capability to analyze unusual poses that have in the past been difficu
lt to reconstruct from single images. To analyze video, we use 3D reconstruction
s from HMR 2.0 as input to a tracking system that operates in 3D. This enables u
s to deal with multiple people and maintain identities through occlusion events.
Our complete approach, 4DHumans, achieves state-of-the-art results for tracking
people from monocular video. Furthermore, we demonstrate the effectiveness of H
MR 2.0 on the downstream task of action recognition, achieving significant impro
vements over previous pose-based action recognition approaches. Our code and mod
els are available on the project website: https://shubham-goel.github.io/4dhuman
s/.
****
Ponder: Point Cloud Pre-training via Neural Rendering
Di Huang, Sida Peng, Tong He, Honghui Yang, Xiaowei Zhou, Wanli Ouyang; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 16089-16098
We propose a novel approach to self-supervised learning of point cloud represent
ations by differentiable neural rendering. Motivated by the fact that informativ
e point cloud features should be able to encode rich geometry and appearance cue
s and render realistic images, we train a point-cloud encoder within a devised p
oint-based neural renderer by comparing the rendered images with real images on
massive RGB-D data. The learned point-cloud encoder can be easily integrated int
o various downstream tasks, including not only high-level tasks like 3D detectio
n and segmentation, but low-level tasks like 3D reconstruction and image synthes
is. Extensive experiments on various tasks demonstrate the superiority of our ap
proach compared to existing pre-training methods.
****
Perpetual Humanoid Control for Real-time Simulated Avatars
Zhengyi Luo, Jinkun Cao, AlexanderWinkler, Kris Kitani, Weipeng Xu; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10
895-10904
We present a physics-based humanoid controller that achieves high-fidelity motio
n imitation and fault-tolerant behavior in the presence of noisy input (e.g. pos
e estimates from video or generated from language) and unexpected falls. Our con

troller scales up to learning ten thousand motion clips without using any external stabilizing forces and learns to naturally recover from fail-state. Given reference motion, our controller can perpetually control simulated avatars without requiring resets. At its core, we propose the progressive multiplicative control policy (PMCP), which dynamically allocates new network capacity to learn harder and harder motion sequences. PMCP allows efficient scaling for learning from large-scale motion databases and adding new tasks, such as fail-state recovery, without catastrophic forgetting. We demonstrate the effectiveness of our controller by using it to imitate noisy poses from video-based pose estimators and language-based motion generators in a live and real-time multi-person avatar use case.
****

HollowNeRF: Pruning Hashgrid-Based NeRFs with Trainable Collision Mitigation
Xiufeng Xie, Riccardo Gherardi, Zhihong Pan, Stephen Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3480-3490
Neural radiance fields (NeRF) have garnered significant attention, with recent works such as Instant-NGP accelerating NeRF training and evaluation through a combination of hashgrid-based positional encoding and neural networks. However, effectively leveraging the spatial sparsity of 3D scenes remains a challenge. To cull away unnecessary regions of the feature grid, existing solutions rely on prior knowledge of object shape or periodically estimate object shape during training by repeated model evaluations, which are costly and wasteful. To address this issue, we propose HollowNeRF, a novel compression solution for hashgrid-based NeRF which automatically sparsifies the feature grid during the training phase. Instead of directly compressing dense features, HollowNeRF trains a coarse 3D saliency mask that guides efficient feature pruning, and employs an alternating direction method of multipliers (ADMM) pruner to sparsify the 3D saliency mask during training. By exploiting the sparsity in the 3D scene to redistribute hash collisions, HollowNeRF improves rendering quality while using a fraction of the parameters of comparable state-of-the-art solutions, leading to a better cost-accuracy trade-off. Our method delivers comparable rendering quality to Instant-NGP, while utilizing just 31% of the parameters. In addition, our solution can achieve a PSNR accuracy gain of up to 1dB using only 56% of the parameters.
****

A Complete Recipe for Diffusion Generative Models
Kushagra Pandey, Stephan Mandt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4261-4272
Score-based Generative Models (SGMs) have demonstrated exceptional synthesis outcomes across various tasks. However, the current design landscape of the forward diffusion process remains largely untapped and often relies on physical heuristics or simplifying assumptions. Utilizing insights from the development of scalable Bayesian posterior samplers, we present a complete recipe for formulating forward processes in SGMs, ensuring convergence to the desired target distribution. Our approach reveals that several existing SGMs can be seen as specific manifestations of our framework. Building upon this method, we introduce Phase Space Langevin Diffusion (PSLD), which relies on score-based modeling within an augmented space enriched by auxiliary variables akin to physical phase space. Empirical results exhibit the superior sample quality and improved speed-quality trade-off of PSLD compared to various competing approaches on established image synthesis benchmarks. Remarkably, PSLD achieves sample quality akin to state-of-the-art SGMs (FID: 2.10 for unconditional CIFAR-10 generation). Lastly, we demonstrate the applicability of PSLD in conditional synthesis using pre-trained score networks, offering an appealing alternative as an SGM backbone for future advancements. Code and model checkpoints can be accessed at https://github.com/mandt-lab/PSLD.
****

The Devil is in the Crack Orientation: A New Perspective for Crack Detection
Zhuangzhuang Chen, Jin Zhang, Zhuonan Lai, Guanming Zhu, Zun Liu, Jie Chen, Jianqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6653-6663
Cracks are usually curve-like structures that are the focus of many computer-vis

ion applications (e.g., road safety inspection and surface inspection of industrial facilities). The existing pixel-based crack segmentation methods rely on time-consuming and costly pixel-level annotations. And the object-based crack detection methods exploit the horizontal box to detect the crack without considering crack orientation, resulting in scale variation and intra-class variation. Considering this, we provide a new perspective for crack detection that models the cracks as a series of sub-cracks with the corresponding orientation. However, the vanilla adaptation of the existing oriented object detection methods to the crack detection tasks will result in limited performance, due to the boundary discontinuity issue and the ambiguities in sub-crack orientation. In this paper, we propose a first-of-its-kind oriented sub-crack detector, dubbed as CrackDet, which is derived from a novel piecewise angle definition, to ease the boundary discontinuity problem. And then, we propose a multi-branch angle regression loss for learning sub-crack orientation and variance together. Since there are no related benchmarks, we construct three fully annotated datasets, namely, ORC, ONPP, and OCCSD, which involve various cracks in road pavement and industrial facilities. Experiments show that our approach outperforms state-of-the-art crack detectors.
****

FedPD: Federated Open Set Recognition with Parameter Disentanglement

Chen Yang, Meilu Zhu, Yifan Liu, Yixuan Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4882-4891

Existing federated learning (FL) approaches are deployed under the unrealistic closed-set setting, with both training and testing classes belong to the same set, which makes the global model fail to identify the unseen classes as `unknown'. To this end, we aim to study a novel problem of federated open-set recognition (FedOSR), which learns an open-set recognition (OSR) model under federated paradigm such that it classifies seen classes while at the same time detects unknown classes. In this work, we propose a parameter disentanglement guided federated open-set recognition (FedPD) algorithm to address two core challenges of FedOSR: cross-client inter-set interference between learning closed-set and open-set knowledge and cross-client intra-set inconsistency by data heterogeneity. The proposed FedPD framework mainly leverages two modules, i.e., local parameter disentanglement (LPD) and global divide-and-conquer aggregation (GDCA), to first disentangle client OSR model into different subnetworks, then align the corresponding parts cross clients for matched model aggregation. Specifically, on the client side, LPD decouples an OSR model into a closed-set subnetwork and an open-set subnetwork by the task-related importance, thus preventing inter-set interference. On the server side, GDCA first partitions the two subnetworks into specific and shared parts, and subsequently aligns the corresponding parts through optimal transport to eliminate parameter misalignment. Extensive experiments on various datasets demonstrate the superior performance of our proposed method.
****

WaterMask: Instance Segmentation for Underwater Imagery

Shijie Lian, Hua Li, Runmin Cong, Suqi Li, Wei Zhang, Sam Kwong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1305-1315

Underwater image instance segmentation is a fundamental and critical step in underwater image analysis and understanding. However, the paucity of general multiclass instance segmentation datasets has impeded the development of instance segmentation studies for underwater images. In this paper, we propose the first underwater image instance segmentation dataset (UIIS), which provides 4628 images for 7 categories with pixel-level annotations. Meanwhile, we also design WaterMask for underwater image instance segmentation for the first time. In Water- Mask, we first devise Difference Similarity Graph Attention Module (DSGAT) to recover lost detailed information due to image quality degradation and downsampling to help the network prediction. Then, we propose Multi-level Feature Refinement Module (MFRM) to predict foreground masks and boundary masks separately by features at different scales, and guide the network through Boundary Mask Strategy (BMS) with boundary learning loss to provide finer prediction results. Extensive experimental results demonstrates that WaterMask can achieve significant gains of 2.9

, 3.8 mAP over Mask R-CNN when using ResNet-50 and ResNet-101. Code and Dataset are available at https: //github.com/LiamLian0727/WaterMask.
****
Score Priors Guided Deep Variational Inference for Unsupervised Real-World Single Image Denoising
Jun Cheng, Tao Liu, Shan Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12937-12948
Real-world single image denoising is crucial and practical in computer vision. Bayesian inversions combined with score priors now have proven effective for single image denoising but are limited to white Gaussian noise. Moreover, applying existing score-based methods for real-world denoising requires not only the explicit train of score priors on the target domain but also the careful design of sampling procedures for posterior inference, which is complicated and impractical. To address these limitations, we propose a score priors-guided deep variational inference, namely ScoreDVI, for practical real-world denoising. By considering the deep variational image posterior with a Gaussian form, score priors are extracted based on easily accessible minimum MSE Non-i.i.d Gaussian denoisers and variational samples, which in turn facilitate optimizing the variational image posterior. Such a procedure adaptively applies cheap score priors to denoising. Additionally, we exploit a Non-i.i.d Gaussian mixture model and variational noise posterior to model the real-world noise. This scheme also enables the pixel-wise fusion of multiple image priors and variational image posteriors. Besides, we develop a noise-aware prior assignment strategy that dynamically adjusts the weight of image priors in the optimization. Our method outperforms other single image-based real-world denoising methods and achieves comparable performance to dataset-based unsupervised methods.
****
L-DAWA: Layer-wise Divergence Aware Weight Aggregation in Federated Self-Supervised Visual Representation Learning
Yasar Abbas Ur Rehman, Yan Gao, Pedro Porto Buarque de Gusmao, Mina Alibeigi, Jiajun Shen, Nicholas D. Lane; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16464-16473
The ubiquity of camera-enabled devices has led to large amounts of unlabeled image data being produced at the edge. The integration of self-supervised learning (SSL) and federated learning (FL) into one coherent system can potentially offer data privacy guarantees while also advancing the quality and robustness of the learned visual representations without needing to move data around. However, client bias and divergence during FL aggregation caused by data heterogeneity limits the performance of learned visual representations on downstream tasks. In this paper, we propose a new aggregation strategy termed Layer-wise Divergence Aware Weight Aggregation (L-DAWA) to mitigate the influence of client bias and divergence during FL aggregation. The proposed method aggregates weights at the layer-level according to the measure of angular divergence between the clients' model and the global model. Extensive experiments with cross-silo and cross-device settings on CIFAR-10/100 and Tiny ImageNet datasets demonstrate that our methods are effective and obtain new SOTA performance on both contrastive and non-contrastive SSL approaches.
****
Improving Transformer-based Image Matching by Cascaded Capturing Spatially Informative Keypoints
Chenjie Cao, Yanwei Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12129-12139
Learning robust local image feature matching is a fundamental low-level vision task, which has been widely explored in the past few years. Recently, detector-free local feature matchers based on transformers have shown promising results, which largely outperform pure Convolutional Neural Network (CNN) based ones. But correlations produced by transformer-based methods are spatially limited to the center of source views' coarse patches, because of the costly attention learning. In this work, we rethink this issue and find that such matching formulation degrades pose estimation, especially for low-resolution images. So we propose a tra

nsformer-based cascade matching model -- Cascade feature Matching TRansformer (CasMTR), to efficiently learn dense feature correlations, which allows us to choose more reliable matching pairs for the relative pose estimation. Instead of re-training a new detector, we use a simple yet effective Non-Maximum Suppression (NMS) post-process to filter keypoints through the confidence map, and largely improve the matching precision. CasMTR achieves state-of-the-art performance in indoor and outdoor pose estimation as well as visual localization. Moreover, thorough ablations show the efficacy of the proposed components and techniques.
****
Controllable Guide-Space for Generalizable Face Forgery Detection
Ying Guo, Cheng Zhen, Pengfei Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20818-20827
Recent studies on face forgery detection have shown satisfactory performance for methods involved in training datasets, but are not ideal enough for unknown domains. This motivates many works to improve the generalization, but forgery-irrelevant information, such as image background and identity, still exists in different domain features and causes unexpected clustering, limiting the generalization.
In this paper, we propose a controllable guide-space (GS) method to enhance the discrimination of different forgery domains, so as to increase the forgery relevance of features and thereby improve the generalization.
The well-designed guide-space can simultaneously achieve both the proper separation of forgery domains and the large distance between real-forgery domains in an explicit and controllable manner. Moreover, for better discrimination, we use a decoupling module to weaken the interference of forgery-irrelevant correlations between domains. Furthermore, we make adjustments to the decision boundary manifold according to the clustering degree of the same domain features within the neighborhood.
Extensive experiments in multiple in-domain and cross-domain settings confirm that our method can achieve state-of-the-art generalization.
****
Calibrating Uncertainty for Semi-Supervised Crowd Counting
Chen LI, Xiaoling Hu, Shahira Abousamra, Chao Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16731-16741
Semi-supervised crowd counting is an important yet challenging task. A popular approach is to iteratively generate pseudo-labels for unlabeled data and add them to the training set. The key is to use uncertainty to select reliable pseudo-labels. In this paper, we propose a novel method to calibrate model uncertainty for crowd counting. Our method takes a supervised uncertainty estimation strategy to train the model through a surrogate function. This ensures the uncertainty is well controlled throughout the training. We propose a matching-based patch-wise surrogate function to better approximate uncertainty for crowd counting tasks. The proposed method pays a sufficient amount of attention to details, while maintaining a proper granularity. Altogether our method is able to generate reliable uncertainty estimation, high quality pseudolabels, and achieve state-of-the-art performance in semisupervised crowd counting.
****
MosaiQ: Quantum Generative Adversarial Networks for Image Generation on NISQ Computers
Daniel Silver, Tirthak Patel, William Cutler, Aditya Ranjan, Harshitta Gandhi, Devesh Tiwari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7030-7039
Quantum machine learning and vision have come to the fore recently, with hardware advances enabling rapid advancement in the capabilities of quantum machines. Recently, quantum image generation has been explored with many potential advantages over non-quantum techniques; however, previous techniques have suffered from poor quality and robustness. To address these problems, we introduce MosaiQ a high-quality quantum image generation GAN framework that can be executed on today's Near-term Intermediate Scale Quantum (NISQ) computers.
****

DVIS: Decoupled Video Instance Segmentation Framework

Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, Pengfei Wan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1282-1291

Video instance segmentation (VIS) is a critical task with diverse applications, including autonomous driving and video editing. Existing methods often underperform on complex and long videos in real world, primarily due to two factors. Firstly, offline methods are limited by the tightly-coupled modeling paradigm, which treats all frames equally and disregards the interdependencies between adjacent frames. Consequently, this leads to the introduction of excessive noise during long-term temporal alignment. Secondly, online methods suffer from inadequate utilization of temporal information. To tackle these challenges, we propose a decoupling strategy for VIS by dividing it into three independent sub-tasks: segmentation, tracking, and refinement. The efficacy of the decoupling strategy relies on two crucial elements: 1) attaining precise long-term alignment outcomes via frame-by-frame association during tracking, and 2) the effective utilization of temporal information predicated on the aforementioned accurate alignment outcomes during refinement. We introduce a novel referring tracker and temporal refiner to construct the Decoupled VIS framework (DVIS). DVIS achieves new SOTA performance in both VIS and VPS, surpassing the current SOTA methods by 7.3 AP and 9.6 VPQ on the OVIS and VIPSeg datasets, which are the most challenging and realistic benchmarks. Moreover, thanks to the decoupling strategy, the referring tracker and temporal refiner are super light-weight (only 6% of the segmenter FLOPs), allowing for efficient training and inference on a single GPU with 11G memory. To promote reproducibility and facilitate further research, we will make the code publicly available.

****

Segmentation of Tubular Structures Using Iterative Training with Tailored Samples

Wei Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23643-23652

We propose a minimal path method to simultaneously compute segmentation masks and extract centerlines of tubular structures with line-topology. Minimal path methods are commonly used for the segmentation of tubular structures in a wide variety of applications. Recent methods use features extracted by CNNs, and often outperform methods using hand-tuned features. However, for CNN-based methods, the samples used for training may be generated inappropriately, so that they can be very different from samples encountered during inference. We approach this discrepancy by introducing a novel iterative training scheme, which enables generating better training samples specifically tailored for the minimal path methods without changing existing annotations. In our method, segmentation masks and centerlines are not determined after one another by post-processing, but obtained using the same steps. Our method requires only very few annotated training images. Comparison with seven previous approaches on three public datasets, including satellite images and medical images, shows that our method achieves state-of-the-art results both for segmentation masks and centerlines.

****

Boundary-Aware Divide and Conquer: A Diffusion-Based Solution for Unsupervised Shadow Removal

Lanqing Guo, Chong Wang, Wenhan Yang, Yufei Wang, Bihan Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13045-13054

Recent deep learning methods have achieved superior results in shadow removal. However, most of these supervised methods rely on training over a huge amount of shadow and shadow-free image pairs, which require laborious annotations and may end up with poor model generalization. Shadows, in fact, only form partial degradation in images, while their non-shadow regions provide rich structural information potentially for unsupervised learning. In this paper, we propose a novel diffusion-based solution for unsupervised shadow removal, which separately models the shadow, non-shadow, and their boundary regions. We employ a pretrained uncon

ditional diffusion model fused with non-corrupted information to generate the natural shadow-free image. While the diffusion model can restore the clear structure in the boundary region by utilizing its adjacent non-corrupted contextual information, it fails to address the inner shadow area due to the isolation of the non-corrupted contexts. Thus we further propose a Shadow-Invariant Intrinsic Decomposition module to exploit the underlying reflectance in the shadow region to maintain structural consistency during the diffusive sampling. Extensive experiments on the publicly available shadow removal datasets show that the proposed method achieves a significant improvement compared to existing unsupervised methods, and even is comparable with some existing supervised methods.
****

Towards Nonlinear-Motion-Aware and Occlusion-Robust Rolling Shutter Correction
Delin Qu, Yizhen Lao, Zhigang Wang, Dong Wang, Bin Zhao, Xuelong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10680-10688
This paper addresses the problem of rolling shutter correction in complex nonlinear and dynamic scenes with extreme occlusion. Existing methods suffer from two main drawbacks. Firstly, they face challenges in estimating the
accurate correction field due to the uniform velocity assumption, leading to significant image correction errors under complex motion. Secondly, the drastic occlusion in dynamic scenes prevents current solutions from achieving better image quality because of the inherent difficulties in aligning and aggregating multiple frames. To tackle these challenges, we model the curvilinear trajectory of pixels analytically and propose a geometry-based Quadratic Rolling Shutter (QRS) motion solver, which precisely estimates the high-order correction field of individual pixels. Besides, to reconstruct high-quality occlusion frames in dynamic scenes, we present a 3D video architecture that effectively Aligns and Aggregates multi-frame context, namely, RSA2-Net. We evaluate our method across a broad range of cameras and video sequences, demonstrating its significant superiority. Specifically, our method surpasses the state-of-the-art by +4.98, +0.77, and +4.33 of PSNR on Carla-RS, Fastec-RS, and BS-RSC datasets, respectively. Code is available at https://github.com/DelinQu/qrsc.
****

Surface Extraction from Neural Unsigned Distance Fields
Congyi Zhang, Guying Lin, Lei Yang, Xin Li, Taku Komura, Scott Schaefer, John Keyser, Wenping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22531-22540
We propose a method, named DualMesh-UDF, to extract a surface from unsigned distance functions (UDFs), encoded by neural networks, or neural UDFs. Neural UDFs are becoming increasingly popular for surface representation because of their versatility in presenting surfaces with arbitrary topologies, as opposed to the signed distance function that is limited to representing a closed surface. However, the applications of neural UDFs are hindered by the notorious difficulty in extracting the target surfaces they represent. Recent methods for surface extraction from a neural UDF suffer from significant geometric errors or topological artifacts due to two main difficulties: (1) A UDF does not exhibit sign changes; and (2) A neural UDF typically has substantial approximation errors.
DualMesh-UDF addresses these two difficulties. Specifically, given a neural UDF encoding a target surface S to be recovered, we first estimate the tangent planes of S at a set of sample points close to S. Next, we organize these sample points into local clusters, and for each local cluster, solve a linear least squares problem to determine a final surface point. These surface points are then connected to create the output mesh surface, which approximates the target surface. The robust estimation of the tangent planes of the target surface and the subsequent minimization problem constitute our core strategy, which contributes to the favorable performance of DualMesh-UDF over other competing methods. To efficiently implement this strategy, we employ an adaptive Octree. Within this framework, we estimate the location of a surface point in each of the octree cells identified as containing part of the target surface. Extensive experiments show that our method outperforms existing methods in terms of surface reconstruction qualit

y while maintaining comparable computational efficiency.
****
CBA: Improving Online Continual Learning via Continual Bias Adaptor
Quanziang Wang, Renzhen Wang, Yichen Wu, Xixi Jia, Deyu Meng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19082-19092
Online continual learning (CL) aims to learn new knowledge and consolidate previously learned knowledge from non-stationary data streams. Due to the time-varying training setting, the model learned from a changing distribution easily forgets the previously learned knowledge and biases towards the newly received task. To address this problem, we propose a Continual Bias Adaptor (CBA) module to augment the classifier network to adapt to catastrophic distribution change during training, such that the classifier network is able to learn a stable consolidation of previously learned tasks. In the testing stage, CBA can be removed which introduces no additional computation cost and memory overhead. We theoretically reveal the reason why the proposed method can effectively alleviate catastrophic distribution shifts, and empirically demonstrate its effectiveness through extensive experiments based on four rehearsal-based baselines and three public continual learning benchmarks.
****
GraphEcho: Graph-Driven Unsupervised Domain Adaptation for Echocardiogram Video Segmentation
Jiewen Yang, Xinpeng Ding, Ziyang Zheng, Xiaowei Xu, Xiaomeng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11878-11887
Echocardiogram video segmentation plays an important role in cardiac disease diagnosis. This paper studies the unsupervised domain adaption (UDA) for echocardiogram video segmentation, where the goal is to generalize the model trained on the source domain to other unlabeled target domains. Existing UDA segmentation methods are not suitable for this task because they do not model local information and the cyclical consistency of heartbeat. In this paper, we introduce a newly collected CardiacUDA dataset and a novel GraphEcho method for cardiac structure segmentation. Our GraphEcho comprises two innovative modules, the Spatial-wise Cross-domain Graph Matching (SCGM) and the Temporal Cycle Consistency (TCC) module, which utilize prior knowledge of echocardiogram videos, i.e., consistent cardiac structure across patients and centers and the heartbeat cyclical consistency, respectively. These two modules can better align global and local features from source and target domains, leading to improved UDA segmentation results. Experimental results showed that our GraphEcho outperforms existing state-of-the-art UDA segmentation methods. Our collected dataset and code will be publicly released upon acceptance. This work will lay a new and solid cornerstone for cardiac structure segmentation from echocardiogram videos.
****
Multi-view Spectral Polarization Propagation for Video Glass Segmentation
Yu Qiao, Bo Dong, Ao Jin, Yu Fu, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, Xin Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23218-23228
In this paper, we present the first polarization-guided video glass segmentation propagation solution (PGVS-Net) that can robustly and coherently propagate glass segmentation in RGB-P video sequences. By leveraging spatiotemporal polarization and color information, our method combines multi-view polarization cues and thus can alleviate the view dependence of single-input intensity variations on glass objects. We demonstrate that our model can outperform glass segmentation on RGB-only video sequences as well as produce more robust segmentation than per-frame RGB-P single-image segmentation methods. To
train and validate PGVS-Net, we introduce a novel RGB-P Glass Video dataset (PGV-117) containing 117 video sequences of scenes captured with different types of camera paths, lighting conditions, dynamics, and glass types.
****
Rethinking Amodal Video Segmentation from Learning Supervised Signals with Objec

t-centric Representation
Ke Fan, Jingshi Lei, Xuelin Qian, Miaopeng Yu, Tianjun Xiao, Tong He, Zheng Zhan g, Yanwei Fu; Proceedings of the IEEE/CVF International Conference on Computer V ision (ICCV), 2023, pp. 1272-1281
Video amodal segmentation is a particularly challenging task in computer vision, which requires to deduce the full shape of an object from the visible parts of it. Recently, some studies have achieved promising performance by using motion f low to integrate information across frames under a self-supervised setting. Howe ver, motion flow has a clear limitation by the two factors of moving cameras and object deformation. This paper presents a rethinking to previous works. We part icularly leverage the supervised signals with object-centric representation in r eal-world scenarios. The underlying idea is the supervision signal of the specif ic object and the features from different views can mutually benefit the deducti on of the full mask in any specific frame. We thus propose an Efficient object-c entric Representation amodal Segmentation (EoRaS). Specially, beyond solely rely ing on supervision signals, we design a translation module to project image feat ures into the Bird's-Eye View (BEV), which introduces 3D information to improve current feature quality. Furthermore, we propose a multi-view fusion layer based temporal module which is equipped with a set of object slots and interacts with features from different views by attention mechanism to fulfill sufficient obje ct representation completion. As a result, the full mask of the object can be de coded from image features updated by object slots. Extensive experiments on both real-world and synthetic benchmarks demonstrate the superiority of our proposed method, achieving state-of-the-art performance. Our code will be released at ht tps://github.com/kfan21/EoRaS.
****
Augmented Box Replay: Overcoming Foreground Shift for Incremental Object Detecti on
Yuyang Liu, Yang Cong, Dipam Goswami, Xialei Liu, Joost van de Weijer; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11367-11377
In incremental learning, replaying stored samples from previous tasks together w ith current task samples is one of the most efficient approaches to address cata strophic forgetting. However, unlike incremental classification, image replay ha s not been successfully applied to incremental object detection (IOD). In this p aper, we identify the overlooked problem of foreground shift as the main reason for this. Foreground shift only occurs when replaying images of previous tasks a nd refers to the fact that their background might contain foreground objects of the current task. To overcome this problem, a novel and efficient Augmented Box Replay (ABR) method is developed that only stores and replays foreground objects and thereby circumvents the foreground shift problem. In addition, we propose a n innovative Attentive RoI Distillation loss that uses spatial attention from re gion-of-interest (RoI) features to constrain current model to focus on the most important information from old model. ABR significantly reduces forgetting of pr evious classes while maintaining high plasticity in current classes. Moreover, i t considerably reduces the storage requirements when compared to standard image replay. Comprehensive experiments on Pascal-VOC and COCO datasets support the st ate-of-the-art performance of our model.
****
Distilled Reverse Attention Network for Open-world Compositional Zero-Shot Learn ing
Yun Li, Zhe Liu, Saurav Jha, Lina Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1782-1791
Open-World Compositional Zero-Shot Learning (OW-CZSL) aims to recognize new comp ositions of seen attributes and objects. In OW-CZSL, methods built on the conven tional closed-world setting degrade severely due to the unconstrained OW test sp ace. While previous works alleviate the issue by pruning compositions according to external knowledge or correlations in seen pairs, they introduce biases that harm the generalization. Some methods thus predict state and object with indepen dently constructed and trained classifiers, ignoring that attributes are highly

context-dependent and visually entangled with objects. In this paper, we propose a novel Distilled Reverse Attention Network to address the challenges. We also model attributes and objects separately but with different motivations, capturing contextuality and locality, respectively. We further design a reverse-and-distill strategy that learns disentangled representations of elementary components in training data supervised by reverse attention and knowledge distillation. We conduct experiments on three datasets and consistently achieve state-of-the-art (SOTA) performance.
****
DandelionNet: Domain Composition with Instance Adaptive Classification for Domain Generalization
Lanqing Hu, Meina Kan, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19050-19059
Domain generalization (DG) attempts to learn a model on source domains that can well generalize to unseen but different domains. The multiple source domains are innately different in distribution but intrinsically related to each other, e.g., from the same label space. To achieve a generalizable feature, most existing methods attempt to reduce the domain discrepancy by either learning domain-invariant feature, or additionally mining domain-specific feature. In the space of these features, the multiple source domains are either tightly aligned or not aligned at all, which both cannot fully take the advantage of complementary information from multiple domains. In order to preserve more complementary information from multiple domains at the meantime of reducing their domain gap, we propose that the multiple domains should not be tightly aligned but composite together, where all domains are pulled closer but still preserve their individuality respectively. This is achieved by using instance-adaptive classifier specified for each instance's classification, where the instance-adaptive classifier is slightly deviated from a universal classifier shared by samples from all domains. This adaptive classifier deviation allows all instances from the same category but different domains to be dispersed around the class center rather than squeezed tightly, leading to better generalization for unseen domain samples. In result, the multiple domains are harmoniously composite centered on a universal core, like a dandelion, so this work is referred to as DandelionNet. Experiments on multiple DG benchmarks demonstrate that the proposed method can learn a model with better generalization and experiments on source free domain adaption also indicate the versatility.
****
TexFusion: Synthesizing 3D Textures with Text-Guided Image Diffusion Models
Tianshi Cao, Karsten Kreis, Sanja Fidler, Nicholas Sharp, Kangxue Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4169-4181
We present TexFusion(Texture Diffusion), a new method to synthesize textures for given 3D geometries, using only large-scale text-guided image diffusion models. In contrast to recent works that leverage 2D text-to-image diffusion models to distill 3D objects using a slow and fragile optimization process, TexFusion introduces a new 3D-consistent generation technique specifically designed for texture synthesis that employs regular diffusion model sampling on different 2D rendered views. Specifically, we leverage latent diffusion models, apply the diffusion model's denoiser on a set of 2D renders of the 3D object, and aggregate the different denoising predictions on a shared latent texture map. Final RGB output textures are produced by optimizing an intermediate neural color field on the decodings of 2D renders of the latent texture. We thoroughly validate TexFusion and show that we can efficiently generate diverse, high quality and globally coherent textures. We achieve state-of-the-art text-guided texture synthesis performance using only image diffusion models, while avoiding the pitfalls of previous distillation-based methods. The text-conditioning offers detailed control and we also do not rely on any ground truth 3D textures for training. This makes our method very versatile and applicable to a broad range of geometries and texture types. We hope that TexFusion will advance AI-based texturing of 3D assets for applications in virtual reality, game design, simulation, and more.

****
Shift from Texture-bias to Shape-bias: Edge Deformation-based Augmentation for R
obust Object Recognition
Xilin He, Qinliang Lin, Cheng Luo, Weicheng Xie, Siyang Song, Feng Liu, Linlin S
hen; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 1526-1535
Recent studies have shown the vulnerability of CNNs under perturbation noises, w
hich is partially caused by the reason that the well-trained CNNs are too biased
 toward the object texture, i.e., they make predictions mainly based on texture
cues. To reduce this texture-bias, current studies resort to learning augmented
samples with heavily perturbed texture to make networks be more biased toward re
latively stable shape cues. However, such methods usually fail to achieve real s
hape-biased networks due to the insufficient diversity of the shape cues. In thi
s paper, we propose to augment the training dataset by generating semantically m
eaningful shapes and samples, via a shape deformation-based online augmentation,
 namely as SDbOA. The samples generated by our SDbOA have two main merits. First
, the augmented samples with more diverse shape variations enable networks to le
arn the shape cues more elaborately, which encourages the network to be shape-bi
ased. Second, semantic-meaningful shape-augmentation samples could be produced b
y jointly regularizing the generator with object texture and edge-guidance soft
constraint, where the edges are represented more robustly with a self informatio
n guided map to better against the noises on them. Extensive experiments under v
arious perturbation noises demonstrate the obvious superiority of our shape-bias
-motivated model over the state of the arts in terms of robustness performance.
Our code is appended in the supplementary material.
****
Lighting Every Darkness in Two Pairs: A Calibration-Free Pipeline for RAW Denois
ing
Xin Jin, Jia-Wen Xiao, Ling-Hao Han, Chunle Guo, Ruixun Zhang, Xialei Liu, Chong
yi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2023, pp. 13275-13284
Calibration-based methods have dominated RAW image denoising under extremely low
-light environments. However, these methods suffer from several main deficiencie
s: 1) the calibration procedure is laborious and time-consuming, 2) denoisers fo
r different cameras are difficult to transfer, and 3) the discrepancy between sy
nthetic noise and real noise is enlarged by high digital gain. To overcome the a
bove shortcomings, we propose a calibration-free pipeline for Lighting Every Dra
kness (LED), regardless of the digital gain or camera sensor. Instead of calibra
ting the noise parameters and training repeatedly, our method could adapt to a t
arget camera only with fewshot paired data and fine-tuning. In addition, well-de
signed structural modification during both stages alleviates the domain gap betw
een synthetic noise and real noise without any extra computational cost. With 2
pairs for each additional digital gain (in total 6 pairs) and 0.5% iterations, o
ur method achieves superior performance over other calibration-based methods.
****
Data-free Knowledge Distillation for Fine-grained Visual Categorization
Renrong Shao, Wei Zhang, Jianhua Yin, Jun Wang; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2023, pp. 1515-1525
Data-free knowledge distillation (DFKD) is a promising approach for addressing i
ssues related to model compression, security privacy, and transmission restricti
ons. Although the existing methods exploiting DFKD have achieved inspiring achie
vements in coarse-grained classification, in practical applications involving fi
ne-grained classification tasks that require more detailed distinctions between
similar categories, sub-optimal results are obtained. To address this issue, we
propose an approach called DFKD-FGVC that extends DFKD to fine-grained vision ca
tegorization (FGVC) tasks. Our approach utilizes an adversarial distillation fra
mework with attention generator, mixed high-order attention distillation, and se
mantic feature contrast learning. Specifically, we introduce a spatial-wise atte
ntion mechanism to the generator to synthesize fine-grained images with more det
ails of discriminative parts. We also utilize the mixed high-order attention mec

hanism to capture complex interactions among parts and the subtle differences am
ong discriminative features of the fine-grained categories, paying attention to
both local features and semantic context relationships. Moreover, we leverage th
e teacher and student models of the distillation framework to contrast high-leve
l semantic feature maps in the hyperspace, comparing variances of different cate
gories. We evaluate our approach on three widely-used FGVC benchmarks (Aircraft,
 Cars196, and CUB200) and demonstrate its superior performance.
****
MotionBERT: A Unified Perspective on Learning Human Motion Representations
Wentao Zhu, Xiaoxuan Ma, Zhaoyang Liu, Libin Liu, Wayne Wu, Yizhou Wang; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, p
p. 15085-15099
We present a unified perspective on tackling various human-centric video tasks b
y learning human motion representations from large-scale and heterogeneous data
resources. Specifically, we propose a pretraining stage in which a motion encode
r is trained to recover the underlying 3D motion from noisy partial 2D observati
ons. The motion representations acquired in this way incorporate geometric, kine
matic, and physical knowledge about human motion, which can be easily transferre
d to multiple downstream tasks. We implement the motion encoder with a Dual-stre
am Spatio-temporal Transformer (DSTformer) neural network. It could capture long
-range spatio-temporal relationships among the skeletal joints comprehensively a
nd adaptively, exemplified by the lowest 3D pose estimation error so far when tr
ained from scratch. Furthermore, our proposed framework achieves state-of-the-ar
t performance on all three downstream tasks by simply finetuning the pretrained
motion encoder with a simple regression head (1-2 layers), which demonstrates th
e versatility of the learned motion representations. Code and models are availab
le at https://motionbert.github.io/
****
PASTA: Proportional Amplitude Spectrum Training Augmentation for Syn-to-Real Dom
ain Generalization
Prithvijit Chattopadhyay, Kartik Sarangmath, Vivek Vijaykumar, Judy Hoffman; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
3, pp. 19288-19300
Synthetic data offers the promise of cheap and bountiful training data for setti
ngs where labeled real-world data is scarce. However, models trained on syntheti
c data significantly underperform when evaluated on real-world data. In this pap
er, we propose Proportional Amplitude Spectrum Training Augmentation (PASTA), a
simple and effective augmentation strategy to improve out-of-the-box synthetic-t
o-real (syn-to-real) generalization performance. PASTA perturbs the amplitude sp
ectra of synthetic images in the Fourier domain to generate augmented views. Spe
cifically, with PASTA we propose a structured perturbation strategy where high-f
requency components are perturbed relatively more than the low-frequency ones. F
or the tasks of semantic segmentation (GTAV-Real), object detection (Sim10K-Real
), and object recognition (VisDA-C Syn-Real), across a total of 5 syn-to-real sh
ifts, we find that PASTA outperforms more complex state-of-the-art generalizatio
n methods while being complementary to the same.
****
EgoPCA: A New Framework for Egocentric Hand-Object Interaction Understanding
Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu, Cewu Lu, Yu-Wing Tai, Chi-Keun
g Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2023, pp. 5273-5284
With the surge in attention to Egocentric Hand-Object Interaction (Ego-HOI), lar
ge-scale datasets such as Ego4D and EPIC-KITCHENS have been proposed. However, m
ost current research is built on resources derived from third-person video actio
n recognition. This inherent domain gap between first- and third-person action v
ideos, which have not been adequately addressed before, makes current Ego-HOI su
boptimal. This paper rethinks and proposes a new framework as an infrastructure
to advance Ego-HOI recognition by Probing, Curation and Adaption (EgoPCA). We co
ntribute comprehensive pre-train sets, balanced test sets and a new baseline, wh
ich are complete with a training-finetuning strategy. With our new framework, we

not only achieve state-of-the-art performance on Ego-HOI benchmarks but also build several new and effective mechanisms and settings to advance further research. We believe our data and the findings will pave a new way for Ego-HOI understanding. Code and data are available at https://mvig-rhos.com/ego_pca.
****
Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image
Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9043-9053
Reconstructing accurate 3D scenes from images is a long-standing vision task. Due to the ill-posedness of the single-image reconstruction problem, most well-established methods are built upon multi-view geometry. State-of-the-art (SOTA) monocular metric depth estimation methods can only handle a single camera model and are unable to perform mixed-data training due to the metric ambiguity. Meanwhile, SOTA monocular methods trained on large mixed datasets achieve zero-shot generalization by learning affine-invariant depths, which cannot recover real-world metrics. In this work, we show that the key to a zero-shot single-view metric depth model lies in the combination of large-scale data training and resolving the metric ambiguity from various camera models. We propose a canonical camera space transformation module, which explicitly addresses the ambiguity problems and can be effortlessly plugged into existing monocular models. Equipped with our module, monocualr models can be stably trained over 8 millions of images with thousands of camera models, resulting in zero-shot generalization to in-the-wild images with unseen camera settings. Experiments demonstrate SOTA performance of our method on 7 zero-shot benchmarks. Our method can recover the metric 3D structure on randomly collected Internet images, enabling plausible single-image metrology. Downstream tasks can also be significantly improved by naively plug-in our model. E.g., our model relieves the scale drift issues of monocular-SLAM (Fig. 1), leading to metric scale high-quality dense mapping.
****
I Can't Believe There's No Images! Learning Visual Tasks Using only Language Supervision
Sophia Gu, Christopher Clark, Aniruddha Kembhavi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2672-2683
Many high-level skills that are required for computer vision tasks, such as parsing questions, comparing and contrasting semantics, and writing descriptions, are also required in other domains such as natural language processing. In this paper, we ask whether it is possible to learn those skills from text data and then transfer them to vision tasks without ever training on visual training data. Key to our approach is exploiting the joint embedding space of contrastively trained vision and language encoders. In practice, there can be systematic differences between embedding spaces for different modalities in contrastive models, and we analyze how these differences affect our approach and study strategies to mitigate this concern. We produce models using only text training data on four representative tasks: image captioning, visual entailment, visual question answering and visual news captioning, and evaluate them on standard benchmarks using images. We find these models perform close to models trained on images, while surpassing prior work for captioning and visual entailment in this text-only setting by over 9 points, and outperforming all prior work on visual news by over 30 points. We also showcase a variety of stylistic image captioning models that are trained using no image data and no human-curated language data, but instead using readily-available text data from books, the web, or language models.
****
Lightweight Image Super-Resolution with Superpixel Token Interaction
Aiping Zhang, Wenqi Ren, Yi Liu, Xiaochun Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12728-12737
Transformer-based methods have demonstrated impressive results on single-image super-resolution (SISR) task. However, self-attention mechanism is computationally expensive when applied to the entire image. As a result, current approaches divide low-resolution input images into small patches, which are processed separat

ely and then fused to generate high-resolution images. Nevertheless, this conventional regular patch division is too coarse and lacks interpretability, resulting in artifacts and non-similar structure interference during attention operations. To address these challenges, we propose a novel super token interaction network (SPIN). Our method employs superpixels to cluster local similar pixels to form the explicable local regions and utilizes intra-superpixel attention to enable local information interaction. It is interpretable because only similar regions complement each other and dissimilar regions are excluded. Moreover, we design a superpixel cross-attention module to facilitate information propagation via the surrogation of superpixels. Extensive experiments demonstrate that the proposed SPIN model performs favorably against the state-of-the-art SR methods in terms of accuracy and lightweight. Code is available at https://github.com/ArcticHare105/SPIN.

****

Feature Prediction Diffusion Model for Video Anomaly Detection

Cheng Yan, Shiyu Zhang, Yang Liu, Guansong Pang, Wenjun Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5527-5537

Anomaly detection in the video is an important research area and a challenging task in real applications. Due to the unavailability of large-scale annotated anomaly events, most existing video anomaly detection (VAD) methods focus on learning the distribution of normal samples to detect the substantially deviated samples as anomalies. To well learn the distribution of normal motion and appearance, many auxiliary networks are employed to extract foreground object or action information. These high-level semantic features effectively filter the noise from the background to decrease its influence on detection models. However, the capability of these extra semantic models heavily affects the performance of the VAD methods. Motivated by the impressive generative and anti-noise capacity of diffusion model (DM), in this work, we introduce a novel DM-based method to predict the features of video frames for anomaly detection. We aim to learn the distribution of normal samples without any extra high-level semantic feature extraction models involved. To this end, we build two denoising diffusion implicit modules to predict and refine the features. The first module concentrates on feature motion learning, while the last focuses on feature appearance learning. To the best of our knowledge, it is the first DM-based method to predict frame features for VAD. The strong capacity of DMs also enables our method to more accurately predict the normal features than non-DM-based feature prediction-based VAD methods. Extensive experiments show that the proposed approach substantially outperforms state-of-the-art competing methods.

****

RANA: Relightable Articulated Neural Avatars

Umar Iqbal, Akin Caliskan, Koki Nagano, Sameh Khamis, Pavlo Molchanov, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23142-23153

We propose RANA, a relightable and articulated neural avatar for the photorealistic synthesis of humans under arbitrary viewpoints, body poses, and lighting. We only require a short video clip of the person to create the avatar and assume no knowledge about the lighting environment. We present a novel framework to model humans while disentangling their geometry, texture, and also lighting environment from monocular RGB videos. To simplify this otherwise ill-posed task we first estimate the coarse geometry and texture of the person via SMPL+D model fitting and then learn an articulated neural representation for photorealistic image generation. RANA first generates the normal and albedo maps of the person in any given target body pose and then uses spherical harmonics lighting to generate the shaded image in the target lighting environment. We also propose to pretrain RANA using synthetic images and demonstrate that it leads to better disentanglement between geometry and texture while also improving robustness to novel body poses. Finally, we also present a new photorealistic synthetic dataset, Relighting Humans, to quantitatively evaluate the performance of the proposed approach.

****

Iterative Denoiser and Noise Estimator for Self-Supervised Image Denoising
Yunhao Zou, Chenggang Yan, Ying Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13265-13274
With the emergence of powerful deep learning tools, more and more effective deep denoisers have advanced the field of image denoising. However, the huge progress made by these learning-based methods severely relies on large-scale and high-quality noisy/clean training pairs, which limits the practicality in real-world scenarios. To overcome this, researchers have been exploring self-supervised approaches that can denoise without paired data. However, the unavailable noise prior and inefficient feature extraction take these methods away from high practicality and precision. In this paper, we propose a Denoise-Corrupt-Denoise pipeline (DCD-Net) for self-supervised image denoising. Specifically, we design an iterative training strategy, which iteratively optimizes the denoiser and noise estimator, and gradually approaches high denoising performances using only single noisy images without any noise prior. The proposed self-supervised image denoising framework provides very competitive results compared with state-of-the-art methods on widely used synthetic and real-world image denoising benchmarks.
****
MasQCLIP for Open-Vocabulary Universal Image Segmentation
Xin Xu, Tianyi Xiong, Zheng Ding, Zhuowen Tu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 887-898
We present a new method for open-vocabulary universal image segmentation, which is capable of performing instance, semantic, and panoptic segmentation under a unified framework. Our approach, called MasQCLIP, seamlessly integrates with a pre-trained CLIP model by utilizing its dense features, thereby circumventing the need for extensive parameter training. MasQCLIP emphasizes two new aspects when building an image segmentation method with a CLIP model: 1) a student-teacher module to deal with masks of the novel (unseen) classes by distilling information from the base (seen) classes; 2) a fine-tuning process to update model parameters for the queries Q within the CLIP model. Thanks to these two simple and intuitive designs, MasQCLIP is able to achieve state-of-the-art performances with a substantial gain over the competing methods by a large margin across all three tasks, including open-vocabulary instance, semantic, and panoptic segmentation. Project page is at https://masqclip.github.io/.
****
Memory-and-Anticipation Transformer for Online Action Understanding
Jiahao Wang, Guo Chen, Yifei Huang, Limin Wang, Tong Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13824-13835
Most existing forecasting systems are memory-based methods, which attempt to mimic human forecasting ability by employing various memory mechanisms and have progressed in temporal modeling for memory dependency. Nevertheless, an obvious weakness of this paradigm is that it can only model limited historical dependence and can not transcend the past. In this paper, we rethink the temporal dependence of event evolution and propose a novel memory-anticipation-based paradigm to model an entire temporal structure, including the past, present, and future. Based on this idea, we present Memory-and-Anticipation Transformer (MAT), a memory-anticipation-based approach, to address the online action detection and anticipation tasks. In addition, owing to the inherent superiority of MAT, it can process online action detection and anticipation tasks in a unified manner. The proposed MAT model is tested on four challenging benchmarks TVSeries, THUMOS'14, HDD, and EPIC-Kitchens-100, for online action detection and anticipation tasks, and it significantly outperforms all existing methods. Code is available at https://github.com/Echo0125/Memory-and-Anticipation-Transformer .
****
Self-similarity Driven Scale-invariant Learning for Weakly Supervised Person Search
Benzhi Wang, Yang Yang, Jinlin Wu, Guo-jun Qi, Zhen Lei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1813-1822
Weakly supervised person search aims to jointly detect and match persons with only bounding box annotations. Existing approaches typically focus on improving th

e features by exploring the relations of persons. However, scale variation probl
em is a more severe obstacle and under-studied that a person often owns images w
ith different scales (resolutions). For one thing, small-scale images contain le
ss information of a person, thus affecting the accuracy of the generated pseudo
labels. For another, different similarities between cross-scale images of a pers
on increase the difficulty of matching. In this paper, we address it by proposin
g a novel one-step framework, named Self-similarity driven Scale-invariant Learn
ing (SSL). Scale invariance can be explored based on the self-similarity prior t
hat it shows the same statistical properties of an image at different scales. To
 this end, we introduce a Multi-scale Exemplar Branch to guide the network in co
ncentrating on the foreground and learning scale-invariant features by hard exem
plars mining. To enhance the discriminative power of the learned features, we fu
rther introduce a dynamic pseudo label prediction that progressively seeks true
labels for training. Experimental results on two standard benchmarks, i.e., PRW
and CUHK-SYSU datasets, demonstrate that the proposed method can solve scale var
iation problem effectively and perform favorably against state-of-the-art method
s. Code is available at https://github.com/Wangbenzhi/SSL.git.
****
MODA: Mapping-Once Audio-driven Portrait Animation with Dual Attentions
Yunfei Liu, Lijian Lin, Fei Yu, Changyin Zhou, Yu Li; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2023, pp. 23020-23029
Audio-driven portrait animation aims to synthesize portrait videos that are cond
itioned by given audio. Animating high-fidelity and multimodal video portraits h
as a variety of applications. Previous methods have attempted to capture differe
nt motion modes and generate high-fidelity portrait videos by training different
 models or sampling signals from given videos. However, lacking correlation lear
ning between lip-sync and other movements (e.g., head pose/eye blinking) usually
 leads to unnatural results. In this paper, we propose a unified system for mult
i-person, diverse, and high-fidelity talking portrait generation. Our method con
tains three stages, i.e., 1) Mapping-Once network with Dual Attentions (MODA) ge
nerates talking representation from given audio. In MODA, we design a dual-atten
tion module to encode accurate mouth movements and diverse modalities. 2) Facial
 composer network generates dense and detailed face landmarks, and 3) temporal-g
uided render syntheses stable videos. Extensive evaluations demonstrate that the
 proposed system produces more natural and realistic video portraits compared to
 previous methods.
****
Realistic Full-Body Tracking from Sparse Observations via Joint-Level Modeling
Xiaozheng Zheng, Zhuo Su, Chao Wen, Zhou Xue, Xiaojie Jin; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14678-14688
To bridge the physical and virtual worlds for rapidly developed VR/AR applicatio
ns, the ability to realistically drive 3D full-body avatars is of great signific
ance. Although real-time body tracking with only the head-mounted displays (HMDs
) and hand controllers is heavily under-constrained, a carefully designed end-to
-end neural network is of great potential to solve the problem by learning from
large-scale motion data. To this end, we propose a two-stage framework that can
obtain accurate and smooth full-body motions with the three tracking signals of
head and hands only. Our framework explicitly models the joint-level features in
 the first stage and utilizes them as spatiotemporal tokens for alternating spat
ial and temporal transformer blocks to capture joint-level correlations in the s
econd stage. Furthermore, we design a set of loss terms to constrain the task of
 a high degree of freedom, such that we can exploit the potential of our joint-l
evel modeling. With extensive experiments on the AMASS motion dataset and real-c
aptured data, we validate the effectiveness of our designs and show our proposed
 method can achieve more accurate and smooth motion compared to existing approac
hes.
****
MetaF2N: Blind Image Super-Resolution by Learning Efficient Model Adaptation fro
m Faces
Zhicun Yin, Ming Liu, Xiaoming Li, Hui Yang, Longan Xiao, Wangmeng Zuo; Proceedi

ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13033-13044
Due to their highly structured characteristics, faces are easier to recover than natural scenes for blind image super-resolution. Therefore, we can extract the degradation representation of an image from the low-quality and recovered face pairs. Using the degradation representation, realistic low-quality images can then be synthesized to fine-tune the super-resolution model for the real-world low-quality image. However, such a procedure is time-consuming and laborious, and the gaps between recovered faces and the ground-truths further increase the optimization uncertainty. To facilitate efficient model adaptation towards image-specific degradations, we propose a method dubbed MetaF2N, which leverages the contained faces to fine-tune model parameters for adapting to the whole natural image in a meta-learning framework. The degradation extraction and low-quality image synthesis steps are thus circumvented in our MetaF2N, and it requires only one fine-tuning step to get decent performance. Considering the gaps between the recovered faces and ground-truths, we further deploy a MaskNet for adaptively predicting loss weights at different positions to reduce the impact of low-confidence areas. To evaluate our proposed MetaF2N, we have collected a real-world low-quality dataset with one or multiple faces in each image, and our MetaF2N achieves superior performance on both synthetic and realworld datasets. Source code, pre-trained models, and collected datasets are available at https://github.com/yinzhicun/MetaF2N.
****
Lighting up NeRF via Unsupervised Decomposition and Enhancement
Haoyuan Wang, Xiaogang Xu, Ke Xu, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12632-12641
Neural Radiance Field (NeRF) is a promising approach for synthesizing novel views, given a set of images and the corresponding camera poses of a scene. However, images photographed from a low-light scene can hardly be used to train a NeRF model to produce high-quality results, due to their low pixel intensities, heavy noise, and color distortion. Combining existing low-light image enhancement methods with NeRF methods also does not work well due to the view inconsistency caused by the individual 2D enhancement process. In this paper, we propose a novel approach, called Low-Light NeRF (or LLNeRF), to enhance the scene representation and synthesize normal-light novel views directly from sRGB low-light images in an unsupervised manner. The core of our approach is a decomposition of radiance field learning, which allows us to enhance the illumination, reduce noise and correct the distorted colors jointly with the NeRF optimization process. Our method is able to produce novel view images with proper lighting and vivid colors and details, given a collection of camera-finished low dynamic range (8-bits/channel) images from a low-light scene. Experiments demonstrate that our method outperforms existing low-light enhancement methods and NeRF methods.
****
ViM: Vision Middleware for Unified Downstream Transferring
Yutong Feng, Biao Gong, Jianwen Jiang, Yiliang Lv, Yujun Shen, Deli Zhao, Jingren Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11696-11707
Foundation models are pre-trained on massive data and transferred to downstream tasks via fine-tuning. This work presents Vision Middleware (ViM), a new learning paradigm that targets unified transferring from a single foundation model to a variety of downstream tasks. ViM consists of a zoo of lightweight plug-in modules, each of which is independently learned on a midstream dataset with a shared frozen backbone. Downstream tasks can then benefit from an adequate aggregation of the module zoo thanks to the rich knowledge inherited from midstream tasks. There are three major advantages of such a design. From the efficiency aspect, the upstream backbone can be trained only once and reused for all downstream tasks without tuning. From the scalability aspect, we can easily append additional modules to ViM with no influence on existing modules. From the performance aspect, ViM can include as many midstream tasks as possible, narrowing the task gap between upstream and downstream. Considering these benefits, we believe that ViM, w

hich the community could maintain and develop together, would serve as a powerful tool to assist foundation models.
****

DIRE for Diffusion-Generated Image Detection

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, Houqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22445-22455

Diffusion models have shown remarkable success in visual synthesis, but have also raised concerns about potential abuse for malicious purposes. In this paper, we seek to build a detector for telling apart real images from diffusion-generated images. We find that existing detectors struggle to detect images generated by diffusion models, even if we include generated images from a specific diffusion model in their training data. To address this issue, we propose a novel image representation called DIffusion Reconstruction Error (DIRE), which measures the error between an input image and its reconstruction counterpart by a pre-trained diffusion model. We observe that diffusion-generated images can be approximately reconstructed by a diffusion model while real images cannot. It provides a hint that DIRE can serve as a bridge to distinguish generated and real images. DIRE provides an effective way to detect images generated by most diffusion models, and it is general for detecting generated images from unseen diffusion models and robust to various perturbations. Furthermore, we establish a comprehensive diffusion-generated benchmark including images generated by eight diffusion models to evaluate the performance of diffusion-generated image detectors. Extensive experiments on our collected benchmark demonstrate that DIRE exhibits superiority over previous generated-image detectors. The code, models, and dataset are available at https://github.com/ZhendongWang6/DIRE.
****

Ord2Seq: Regarding Ordinal Regression as Label Sequence Prediction

Jinhong Wang, Yi Cheng, Jintai Chen, TingTing Chen, Danny Chen, Jian Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5865-5875

Ordinal regression refers to classifying object instances into ordinal categories. It has been widely studied in many scenarios, such as medical disease grading and movie rating. Known methods focused only on learning inter-class ordinal relationships, but still incur limitations in distinguishing adjacent categories thus far. In this paper, we propose a simple sequence prediction framework for ordinal regression called Ord2Seq, which, for the first time, transforms each ordinal category label into a special label sequence and thus regards an ordinal regression task as a sequence prediction process. In this way, we decompose an ordinal regression task into a series of recursive binary classification steps, so as to subtly distinguish adjacent categories. Comprehensive experiments show the effectiveness of distinguishing adjacent categories for performance improvement and our new approach exceeds state-of-the-art performances in four different scenarios. Codes are available at https://github.com/wjh892521292/Ord2Seq.
****

Bring Clipart to Life

Nanxuan Zhao, Shengqi Dang, Hexun Lin, Yang Shi, Nan Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23341-23350

The development of face editing has been boosted since the birth of StyleGAN. While previous works have explored different interactive methods, such as sketching and exemplar photos, they have been limited in terms of expressiveness and generality. In this paper, we propose a new interaction method by guiding the editing with abstract clipart, composed of a set of simple semantic parts, allowing users to control across face photos with simple clicks. However, this is a challenging task given the large domain gap between colorful face photos and abstract clipart with limited data. To solve this problem, we introduce a framework called ClipFaceShop built on top of StyleGAN. The key idea is to take advantage of W+ latent code encoded rich and disentangled visual features, and create a new lightweight selective feature adaptor to predict a modifiable path toward the target output photo. Since no pairwise labeled data exists for training, we design a

set of losses to provide supervision signals for learning the modifiable path. Experimental results show that ClipFaceShop generates realistic and faithful face photos, sharing the same facial attributes as the reference clipart. We demonstrate that ClipFaceShop supports clipart in diverse styles, even in form of a free-hand sketch.
****
Co-Evolution of Pose and Mesh for 3D Human Body Estimation from Video
Yingxuan You, Hong Liu, Ti Wang, Wenhao Li, Runwei Ding, Xia Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14963-14973
Despite significant progress in single image-based 3D human mesh recovery, accurately and smoothly recovering 3D human motion from a video remains challenging. Existing video-based methods generally recover human mesh by estimating the complex pose and shape parameters from coupled image features, whose high complexity and low representation ability often result in inconsistent pose motion and limited shape patterns. To alleviate this issue, we introduce 3D pose as the intermediary and propose a Pose and Mesh Co-Evolution network (PMCE) that decouples this task into two parts: 1) video-based 3D human pose estimation and 2) mesh vertices regression from the estimated 3D pose and temporal image feature. Specifically, we propose a two-stream encoder that estimates mid-frame 3D pose and extracts a temporal image feature from the input image sequence. In addition, we design a co-evolution decoder that performs pose and mesh interactions with the image-guided Adaptive Layer Normalization (AdaLN) to make pose and mesh fit the human body shape. Extensive experiments demonstrate that the proposed PMCE outperforms previous state-of-the-art methods in terms of both per-frame accuracy and temporal consistency on three benchmark datasets: 3DPW, Human3.6M, and MPI-INF-3DHP. Our code is available at https://github.com/kasvii/PMCE.
****
Noise2Info: Noisy Image to Information of Noise for Self-Supervised Image Denoising
Jiachuan Wang, Shimin Di, Lei Chen, Charles Wang Wai Ng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16034-16043
Unsupervised image denoising has been proposed to alleviate the widespread noise problem without requiring clean images. Existing works mainly follow the self-supervised way, which tries to reconstruct each pixel x of noisy images without the knowledge of x. More recently, some pioneer works further emphasize the importance of x and propose to weigh the information extracted from x and other pixels when recovering x. However, such a method is highly sensitive to the standard deviation \sigma_n of noises injected to clean images, where \sigma_n is inaccessible without knowing clean images. Thus, it is unrealistic to assume that \sigma_n is known for pursuing high model performance.
To alleviate this issue, we propose Noise2Info to extract the critical information, the standard deviation \sigma_n of injected noise, only based on the noisy images. Specifically, we first theoretically provide an upper bound on \sigma_n, while the bound requires clean images. Then, we propose a novel method to estimate the bound of \sigma_n by only using noisy images. Besides, we prove that the difference between our estimation with the true deviation goes smaller as the model training. Empirical studies show that Noise2Info is effective and robust on benchmark data sets and closely estimates the standard deviation of noises during model training.
****
Controllable Visual-Tactile Synthesis
Ruihan Gao, Wenzhen Yuan, Jun-Yan Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7040-7052
Deep generative models have various content creation applications such as graphic design, e-commerce, and virtual try-on. However, current works mainly focus on synthesizing realistic visual outputs, often ignoring other sensory modalities, such as touch, which limits physical interaction with users. In this work, we leverage deep generative models to create a multi-sensory experience where users can touch and see the synthesized object when sliding their fingers on a haptic

surface. The main challenges lie in the significant scale discrepancy between vision and touch sensing and the lack of explicit mapping from touch sensing data to a haptic rendering device. To bridge this gap, we collect high-resolution tactile data with a GelSight sensor and create a new visuotactile clothing dataset. We then develop a conditional generative model that synthesizes both visual and tactile outputs from a single sketch. We evaluate our method regarding image quality and tactile rendering accuracy. Finally, we introduce a pipeline to render high-quality visual and tactile outputs on an electroadhesion-based haptic device for an immersive experience, allowing for challenging materials and editable sketch inputs.
****
Keep It SimPool: Who Said Supervised Transformers Suffer from Attention Deficit?
Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, Yannis Avrithis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5350-5360
Convolutional networks and vision transformers have different forms of pairwise interactions, pooling across layers and pooling at the end of the network. Does the latter really need to be different?
As a by-product of pooling, vision transformers provide spatial attention for free, but this is most often of low quality unless self-supervised, which is not well studied. Is supervision really the problem?

In this work, we develop a generic pooling framework and then we formulate a number of existing methods as instantiations. By discussing the properties of each group of methods, we derive SimPool, a simple attention-based pooling mechanism as a replacement of the default one for both convolutional and transformer encoders. We find that, whether supervised or self-supervised, this improves performance on pre-training and downstream tasks and provides attention maps delineating object boundaries in all cases. One could thus call SimPool universal. To our knowledge, we are the first to obtain attention maps in supervised transformers of at least as good quality as self-supervised, without explicit losses or modifying the architecture. Code at: https://github.com/billpsomas/simpool.
****
SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling
Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, Lei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20282-20292
Synthetic data has emerged as a promising source for 3D human research as it offers low-cost access to large-scale human datasets. To advance the diversity and annotation quality of human models, we introduce a new synthetic dataset, SynBody, with three appealing features: 1) a clothed parametric human model that can generate a diverse range of subjects; 2) the layered human representation that naturally offers high-quality 3D annotations to support multiple tasks; 3) a scalable system for producing realistic data to facilitate real-world tasks. The dataset comprises 1.2M images with corresponding accurate 3D annotations, covering 10,000 human body models, 1,187 actions, and various viewpoints. The dataset includes two subsets for human pose and shape estimation as well as human neural rendering. Extensive experiments on SynBody indicate that it substantially enhances both SMPL and SMPL-X estimation. Furthermore, the incorporation of layered annotations offers a valuable training resource for investigating the Human Neural Radiance Fields(NeRF).
****
Viewset Diffusion: (0-)Image-Conditioned 3D Generative Models from 2D Data
Stanislaw Szymanowicz, Christian Rupprecht, Andrea Vedaldi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8863-8873
We present Viewset Diffusion, a diffusion-based generator that outputs 3D objects while only using multi-view 2D data for supervision. We note that there exists a one-to-one mapping between viewsets, i.e., collections of several 2D views of

an object, and 3D models. Hence, we train a diffusion model to generate viewsets, but design the neural network generator to reconstruct internally corresponding 3D models, thus generating those too. We fit a diffusion model to a large number of viewsets for a given category of objects. The resulting generator can be conditioned on zero, one or more input views. Conditioned on a single view, it performs 3D reconstruction accounting for the ambiguity of the task and allowing to sample multiple solutions compatible with the input. The model performs reconstruction efficiently, in a feed-forward manner, and is trained using only rendering losses using as few as three views per viewset. Project page: szymanowiczs.github.io/viewset-diffusion
****
LoGoPrompt: Synthetic Text Images Can Be Good Visual Prompts for Vision-Language Models
Cheng Shi, Sibei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2932-2941
Prompt engineering is a powerful tool used to enhance the performance of pre-trained models on downstream tasks. For example, providing the prompt "Let's think step by step" improved GPT-3's reasoning accuracy to 63% on MutiArith while prompting "a photo of" filled with a class name enables CLIP to achieve 80% zero-shot accuracy on ImageNet. While previous research has explored prompt learning for the visual modality, analyzing what constitutes a good visual prompt specifically for image recognition is limited. In addition, existing visual prompt tuning methods' generalization ability is worse than text-only prompting tuning. This paper explores our key insight: synthetic text images are good visual prompts for vision-language models! To achieve that, we propose our LoGoPrompt, which reformulates the classification objective to the visual prompt selection and addresses the chicken-and-egg challenge of first adding synthetic text images as class-wise visual prompts or predicting the class first. Without any trainable visual prompt parameters, experimental results on 16 datasets demonstrate that our method consistently outperforms state-of-the-art methods in few-shot learning, base-to-new generalization, and domain generalization. The code will be publicly available upon publication.
****
EP2P-Loc: End-to-End 3D Point to 2D Pixel Localization for Large-Scale Visual Localization
Minjung Kim, Junseo Koo, Gunhee Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21527-21537
Visual localization is the task of estimating a 6-DoF camera pose of a query image within a provided 3D reference map. Thanks to recent advances in various 3D sensors, 3D point clouds are becoming a more accurate and affordable option for building the reference map, but research to match the points of 3D point clouds with pixels in 2D images for visual localization remains challenging. Existing approaches that jointly learn 2D-3D feature matching suffer from low inliers due to representational differences between the two modalities, and the methods that bypass this problem into classification have an issue of poor refinement. In this work, we propose EP2P-Loc, a novel large-scale visual localization method that mitigates such appearance discrepancy and enables end-to-end training for pose estimation. To increase the number of inliers, we propose a simple algorithm to remove invisible 3D points in the image, and find all 2D-3D correspondences without keypoint detection. To reduce memory usage and search complexity, we take a coarse-to-fine approach where we extract patch-level features from 2D images, then perform 2D patch classification on each 3D point, and obtain the exact corresponding 2D pixel coordinates through positional encoding. Finally, for the first time in this task, we employ a differentiable PnP for end-to-end training. In the experiments on newly curated large-scale indoor and outdoor benchmarks based on 2D-3D-S and KITTI, we show that our method achieves the state-of-the-art performance compared to existing visual localization and image-to-point cloud registration methods.
****
SIRA-PCR: Sim-to-Real Adaptation for 3D Point Cloud Registration

Suyi Chen, Hao Xu, Ru Li, Guanghui Liu, Chi-Wing Fu, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14394-14405

Point cloud registration is essential for many applications. However, existing real datasets require extremely tedious and costly annotations, yet may not provide accurate camera poses. For the synthetic datasets, they are mainly object-level, so the trained models may not generalize well to real scenes. We design SIRA-PCR, a new approach to 3D point cloud registration. First, we build a synthetic scene-level 3D registration dataset, specifically designed with physically-based and random strategies to arrange diverse objects. Second, we account for variations in different sensing mechanisms and layout placements, then formulate a sim-to-real adaptation framework with an adaptive re-sample module to simulate patterns in real point clouds. To our best knowledge, this is the first work that explores sim-to-real adaptation for point cloud registration. Extensive experiments show the SOTA performance of SIRA-PCR on widely-used indoor and outdoor datasets. The code and dataset will be released on https://github.com/Chen-Suyi/SIRA_Pytorch.

****

FeatEnHancer: Enhancing Hierarchical Features for Object Detection and Beyond Under Low-Light Vision

Khurram Azeem Hashmi, Goutham Kallempudi, Didier Stricker, Muhammad Zeshan Afzal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6725-6735

Extracting useful visual cues for the downstream tasks is especially challenging under low-light vision. Prior works create enhanced representations by either correlating visual quality with machine perception or designing illumination-degrading transformation methods that require pre-training on synthetic datasets. We argue that optimizing enhanced image representation pertaining to the loss of the downstream task can result in more expressive representations. Therefore, in this work, we propose a novel module, FeatEnHancer, that hierarchically combines multiscale features using multiheaded attention guided by task-related loss function to create suitable representations. Furthermore, our intra-scale enhancement improves the quality of features extracted at each scale or level, as well as combines features from different scales in a way that reflects their relative importance for the task at hand. FeatEnHancer is a general-purpose plug-and-play module and can be incorporated into any low-light vision pipeline. We show with extensive experimentation that the enhanced representation produced with FeatEnHancer significantly and consistently improves results in several low-light vision tasks, including dark object detection (+5.7 mAP on ExDark), face detection (+1.5 mAP on DARK FACE), nighttime semantic segmentation (+5.1 mIoU on ACDC), and video object detection (+1.8 mAP on DarkVision), highlighting the effectiveness of enhancing hierarchical features under low-light vision.

****

SOAR: Scene-debiasing Open-set Action Recognition

Yuanhao Zhai, Ziyi Liu, Zhenyu Wu, Yi Wu, Chunluan Zhou, David Doermann, Junsong Yuan, Gang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10244-10254

Deep models have the risk of utilizing spurious clues to make predictions, e.g., recognizing actions via classifying the background scene. This problem severely degrades the open-set action recognition performance when the testing samples exhibit scene distributions different from the training samples. To mitigate this scene bias, we propose a Scene-debiasing Open-set Action Recognition method (SOAR), which features an adversarial reconstruction module and an adaptive adversarial scene classification module. The former prevents a decoder from reconstructing the video background given video features, and thus helps reduce the background information in feature learning. The latter aims to confuse scene type classification given video features, and helps to learn scene-invariant information. In addition, we design an experiment to quantify the scene bias. The results suggest current open-set action recognizers are biased toward the scene, and our SO

AR better mitigates such bias. Furthermore, extensive experiments show our metho
d outperforms state-of-the-art methods, with ablation studies demonstrating the
effectiveness of our proposed modules.
****
Physics-Augmented Autoencoder for 3D Skeleton-Based Gait Recognition
Hongji Guo, Qiang Ji; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2023, pp. 19627-19638
In this paper, we introduce physics-augmented autoencoder (PAA), a framework for
 3D skeleton-based human gait recognition. Specifically, we construct the autoen
coder with a graph-convolution-based encoder and a physics-based decoder. The en
coder takes the skeleton sequence as input and generates the generalized positio
ns and forces of each joint, which are taken by the decoder to reconstruct the i
nput skeleton based on the Lagrangian dynamics. In this way, the intermediate re
presentations are physically plausible and discriminative. During the inference,
 the decoder is discared and a RNN-based classifier takes the output of the enco
der for gait recognition. We evaluated our proposed method on three benchmark da
tasets including Gait3D, GREW, and KinectGait. Our method achieves state-of-the-
art performance for 3D skeleton-based gait recognition. Furthermore, extensive a
blation studies show that our method generalizes better and is more robust with
small-scale training data by incorporating the physics knowledge. We also valida
ted the physical plausibility of the intermediate representations by making forc
e predictions on real data with physical annotations.
****
Regularized Primitive Graph Learning for Unified Vector Mapping
Lei Wang, Min Dai, Jianan He, Jingwei Huang; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2023, pp. 16817-16826
Large-scale vector mapping is the foundation for transportation and urban planni
ng. Most existing mapping methods are tailored to one specific mapping task, due
 to task-specific requirements on shape regularization and topology reconstructi
on. We propose GraphMapper, a unified framework for end-to-end vector map extrac
tion from satellite images. Our key idea is using primitive graph as a unified r
epresentation of vector maps and formulating shape regularization and topology r
econstruction as primitive graph reconstruction problems that can be solved in t
he same framework. Specifically, shape regularization is modeled as the consiste
ncy between primitive directions and their pairwise relationship. Based on the p
rimitive graph, we design a learning approach to reconstruct primitive graphs in
 multiple stages. GraphMapper can fully explore primitive-wise and pairwise info
rmation for shape regularization and topology reconstruction, resulting improved
 primitive graph learning capabilities. We empirically demonstrate the effective
ness of GraphMapper on two challenging mapping tasks for building footprints and
 road networks. With the premise of sharing the majority design of the architect
ure and a few task-specific designs, our model outperforms state-of-the-art meth
ods in both tasks on public benchmarks. Our code will be publicly available.
****
Saliency Regularization for Self-Training with Partial Annotations
Shouwen Wang, Qian Wan, Xiang Xiang, Zhigang Zeng; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 1611-1620
Partially annotated images are easy to obtain in multi-label classification. How
ever, unknown labels in partially annotated images exacerbate the positive-negat
ive imbalance inherent in multi-label classification, which affects supervised l
earning of known labels. Most current methods require sufficient image annotatio
ns, and do not focus on the imbalance of the labels in the supervised training p
hase. In this paper, we propose saliency regularization (SR) for a novel self-tr
aining framework. In particular, we model saliency on the class-specific maps, a
nd strengthen the saliency of object regions corresponding to the present labels
. Besides, we introduce consistency regularization to mine unlabeled information
 to complement unknown labels with the help of SR. It is verified to alleviate t
he negative dominance caused by the imbalance, and achieve state-of-the-art perf
ormance on Pascal VOC 2007, MS-COCO, VG-200, and OpenImages V3.
****

Stabilizing Visual Reinforcement Learning via Asymmetric Interactive Cooperation
Yunpeng Zhai, Peixi Peng, Yifan Zhao, Yangru Huang, Yonghong Tian; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 207
-216
Vision-based reinforcement learning (RL) depends on discriminative representatio
n encoders to abstract the observation states. Despite the great success of incr
easing CNN parameters for many supervised computer vision tasks, reinforcement l
earning with temporal-difference (TD) losses cannot benefit from it in most comp
lex environments. In this paper, we analyze that the training instability arises
 from the oscillating self-overfitting of the heavy-optimizable encoder. We argu
e that serious oscillation will occur to the parameters when enforced to fit the
 sensitive TD targets, causing uncertain drifting of the latent state space and
thus transmitting these perturbations to the policy learning. To alleviate this
phenomenon, we propose a novel asymmetric interactive cooperation approach with
the interaction between a heavy-optimizable encoder and a supportive light-optim
izable encoder, in which both their advantages are integrated including the high
ly discriminative capability as well as the training stability. We also present
a greedy bootstrapping optimization to isolate the visual perturbations from pol
icy learning, where representation and policy are trained sufficiently by turns.
 Finally, we demonstrate the effectiveness of our method in utilizing larger vis
ual models by first-person highway driving task CARLA and Vizdoom environments.
****
FlipNeRF: Flipped Reflection Rays for Few-shot Novel View Synthesis
Seunghyeon Seo, Yeonjin Chang, Nojun Kwak; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2023, pp. 22883-22893
Neural Radiance Field (NeRF) has been a mainstream in novel view synthesis with
its remarkable quality of rendered images and simple architecture. Although NeRF
 has been developed in various directions improving continuously its performance
, the necessity of a dense set of multi-view images still exists as a stumbling
block to progress for practical application. In this work, we propose FlipNeRF,
a novel regularization method for few-shot novel view synthesis by utilizing our
 proposed flipped reflection rays. The flipped reflection rays are explicitly de
rived from the input ray directions and estimated normal vectors, and play a rol
e of effective additional training rays while enabling to estimate more accurate
 surface normals and learn the 3D geometry effectively. Since the surface normal
 and the scene depth are both derived from the estimated densities along a ray,
the accurate surface normal leads to more exact depth estimation, which is a key
 factor for few-shot novel view synthesis. Furthermore, with our proposed Uncert
ainty-aware Emptiness Loss and Bottleneck Feature Consistency Loss, FlipNeRF is
able to estimate more reliable outputs with reducing floating artifacts effectiv
ely across the different scene structures, and enhance the feature-level consist
ency between the pair of the rays cast toward the photo-consistent pixels withou
t any additional feature extractor, respectively. Our FlipNeRF achieves the SOTA
 performance on the multiple benchmarks across all the scenarios.
****
Discovering Spatio-Temporal Rationales for Video Question Answering
Yicong Li, Junbin Xiao, Chun Feng, Xiang Wang, Tat-Seng Chua; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13869-13
878
This paper strives to solve complex video question answering (VideoQA) which fea
tures long videos containing multiple objects and events at different time. To t
ackle the challenge, we highlight the importance of identifying question-critica
l temporal moments and spatial objects from the vast amount of video content. To
wards this, we propose a Spatio-Temporal Rationalizer (STR), a differentiable se
lection module that adaptively collects question-critical moments and objects us
ing cross-modal interaction. The discovered video moments and objects are then s
erved as grounded rationales to support answer reasoning. Based on STR, we furth
er propose TranSTR, a Transformer-style neural network architecture that takes S
TR as the core and additionally underscores a novel answer interaction mechanism
 to coordinate STR for answer decoding. Experiments on four datasets show that T

ranSTR achieves new state-of-the-art (SoTA). Especially, on NExT-QA and Causal-VidQA which feature complex VideoQA, it significantly surpasses the previous SoTA by 5.8% and 6.8%, respectively. We then conduct extensive studies to verify the importance of STR as well as the proposed answer interaction mechanism. With the success of TranSTR and our comprehensive analysis, we hope this work can spark more future efforts in complex VideoQA. Our results are fully reproducible at https://anonymous.4open.science/r/TranSTR/.
****
Iterative Soft Shrinkage Learning for Efficient Image Super-Resolution
Jiamian Wang, Huan Wang, Yulun Zhang, Yun Fu, Zhiqiang Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12590-12599
Image super-resolution (SR) has witnessed extensive neural network designs from CNN to transformer architectures. However, prevailing SR models suffer from prohibitive memory footprint and intensive computations, which limits further deployment on edge devices. This work investigates the potential of network pruning for super-resolution to take advantage of off-the-shelf network designs and reduce the underlying computational overhead. Two main challenges remain in applying pruning methods for SR. First, the widely-used filter pruning technique reflects limited granularity and restricted adaptability to diverse network structures. Second, existing pruning methods generally operate upon a pre-trained network for the sparse structure determination, hard to get rid of dense model training in the traditional SR paradigm. To address these challenges, we adopt unstructured pruning with sparse models directly trained from scratch. Specifically, we propose a novel Iterative Soft Shrinkage-Percentage (ISS-P) method by optimizing the sparse structure of a randomly initialized network at each iteration and tweaking unimportant weights with a small amount proportional to the magnitude scale on-the-fly. We observe that the proposed ISS-P can dynamically learn sparse structures adapting to the optimization process and preserve the sparse model's trainability by yielding a more regularized gradient throughput. Experiments on benchmark datasets demonstrate the effectiveness of the proposed ISS-P over diverse network architectures. Code is available at https://github.com/Jiamian-Wang/Iterative-Soft-Shrinkage-SR
****
Learning Hierarchical Features with Joint Latent Space Energy-Based Prior
Jiali Cui, Ying Nian Wu, Tian Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2218-2227
This paper studies the fundamental problem of multi-layer generator models in learning hierarchical representations. The multi-layer generator model that consists of multiple layers of latent variables organized in a top-down architecture tends to learn multiple levels of data abstraction. However, such multi-layer latent variables are typically parameterized to be Gaussian, which can be less informative in capturing complex abstractions, resulting in limited success in hierarchical representation learning. On the other hand, the energy-based (EBM) prior is known to be expressive in capturing the data regularities, but it often lacks the hierarchical structure to capture different levels of hierarchical representations. In this paper, we propose a joint latent space EBM prior model with multi-layer latent variables for effective hierarchical representation learning. We develop a variational joint learning scheme that seamlessly integrates an inference model for efficient inference. Our experiments demonstrate that the proposed joint EBM prior is effective and expressive in capturing hierarchical representations and modelling data distribution.
****
UniFormerV2: Unlocking the Potential of Image ViTs for Video Understanding
Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Limin Wang, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1632-1643
The prolific performances of Vision Transformers (ViTs) in image tasks have prompted research into adapting the image ViTs for video tasks. However, the substantial gap between image and video impedes the spatiotemporal learning of these im

age-pretrained models. Though video-specialized models like UniFormer can transfer to the video domain more seamlessly, their unique architectures require prolonged image pretraining, limiting the scalability. Given the emergence of powerful open-source image ViTs, we propose unlocking their potential for video understanding with efficient UniFormer designs. We call the resulting model UniFormerV2, since it inherits the concise style of the UniFormer block, while redesigning local and global relation aggregators that seamlessly integrate advantages from both ViTs and UniFormer. Our UniFormerV2 achieves state-of-the-art performances on 8 popular video benchmarks, including scene-related Kinetics-400/600/700, heterogeneous Moments in Time, temporal-related Something-Something V1/V2, and untrimmed ActivityNet and HACS. It is noteworthy that to the best of our knowledge, UniFormerV2 is the first to elicit 90% top-1 accuracy on Kinetics-400.
****
G2L: Semantically Aligned and Uniform Video Grounding via Geodesic and Game Theory
Hongxiang Li, Meng Cao, Xuxin Cheng, Yaowei Li, Zhihong Zhu, Yuexian Zou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12032-12042
The recent video grounding works attempt to introduce vanilla contrastive learning into video grounding. However, we claim that this naive solution is suboptimal. Contrastive learning requires two key properties: (1) alignment of features of similar samples, and (2) uniformity of the induced distribution of the normalized features on the hypersphere. Due to two annoying issues in video grounding: (1) the co-existence of some visual entities in both ground truth and other moments, i.e. semantic overlapping; (2) only a few moments in the video are annotated, i.e. sparse annotation dilemma, vanilla contrastive learning is unable to model the correlations between temporally distant moments and learned inconsistent video representations. Both characteristics lead to vanilla contrastive learning being unsuitable for video grounding. In this paper, we introduce Geodesic and Game Localization (G2L), a semantically aligned and uniform video grounding framework via geodesic and game theory. We quantify the correlations among moments leveraging the geodesic distance that guides the model to learn the correct cross-modal representations. Furthermore, from the novel perspective of game theory, we propose semantic Shapley interaction based on geodesic distance sampling to learn fine-grained semantic alignment in similar moments. Experiments on three benchmarks demonstrate the effectiveness of our method.
****
TARGET: Federated Class-Continual Learning via Exemplar-Free Distillation
Jie Zhang, Chen Chen, Weiming Zhuang, Lingjuan Lyu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4782-4793
This paper focuses on an under-explored yet important problem: Federated Class-Continual Learning (FCCL), where new classes are dynamically added in federated learning. Existing FCCL works suffer from various limitations, such as requiring additional datasets or storing the private data from previous tasks. In response, we first demonstrate that non-IID data exacerbates catastrophic forgetting issue in FL. Then we propose a novel method called TARGET (federatTed clAss-continual leaRninG via Exemplar-free disTillation), which alleviates catastrophic forgetting in FCCL while preserving client data privacy. Our proposed method leverages the previously trained global model to transfer knowledge of old tasks to the current task at the model level. Moreover, a generator is trained to produce synthetic data to simulate the global distribution of data on each client at the data level. Compared to previous FCCL methods, TARGET does not require any additional datasets or storing real data from previous tasks, which makes it ideal for data-sensitive scenarios.
****
FashionNTM: Multi-turn Fashion Image Retrieval via Cascaded Memory
Anwesan Pal, Sahil Wadhwa, Ayush Jaiswal, Xu Zhang, Yue Wu, Rakesh Chada, Pradeep Natarajan, Henrik I. Christensen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11323-11334
Multi-turn textual feedback-based fashion image retrieval focuses on a real-worl

d setting, where users can iteratively provide information to refine retrieval r
esults until they find an item that fits all their requirements. In this work, w
e present a novel memory-based method, called FashionNTM, for such a multi-turn
system. Our framework incorporates a new Cascaded Memory Neural Turing Machine (
CM-NTM) approach for implicit state management, thereby learning to integrate in
formation across all past turns to retrieve new images, for a given turn. Unlike
 vanilla Neural Turing Machine (NTM), our CM-NTM operates on multiple inputs, wh
ich interact with their respective memories via individual read and write heads,
 to learn complex relationships. Extensive evaluation results show that our prop
osed method outperforms the previous state-of-the-art algorithm by 50.5%, on Mul
ti-turn FashionIQ -- the only existing multi-turn fashion dataset currently, in
addition to having a relative improvement of 12.6% on Multi-turn Shoes -- an ext
ension of the single-turn Shoes dataset that we created in this work. Further an
alysis of the model in a real-world interactive setting demonstrates two importa
nt capabilities of our model -- memory retention across turns, and agnosticity t
o turn order for non-contradictory feedback. Finally, user study results show th
at images retrieved by FashionNTM were favored by 83.1% over other multi-turn mo
dels.
****
MolGrapher: Graph-based Visual Recognition of Chemical Structures
Lucas Morin, Martin Danelljan, Maria Isabel Agea, Ahmed Nassar, Valery Weber, In
gmar Meijer, Peter Staar, Fisher Yu; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 19552-19561
The automatic analysis of chemical literature has immense potential to accelerat
e the discovery of new materials and drugs. Much of the critical information in
patent documents and scientific articles is contained in figures, depicting the
molecule structures. However, automatically parsing the exact chemical structure
 is a formidable challenge, due to the amount of detailed information, the diver
sity of drawing styles, and the need for training data. In this work, we introdu
ce MolGrapher to recognize chemical structures visually. First, a deep keypoint
detector detects the atoms. Second, we treat all candidate atoms and bonds as no
des and put them in a graph. This construct allows a natural graph representatio
n of the molecule. Last, we classify atom and bond nodes in the graph with a Gra
ph Neural Network. To address the lack of real training data, we propose a synth
etic data generation pipeline producing diverse and realistic results. In additi
on, we introduce a large-scale benchmark of annotated real molecule images, USPT
O-30K, to spur research on this critical topic. Extensive experiments on five da
tasets show that our approach significantly outperforms classical and learning-b
ased methods in most settings. Code, models, and datasets are available.
****
SAMPLING: Scene-adaptive Hierarchical Multiplane Images Representation for Novel
 View Synthesis from a Single Image
Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, Ming-Hsuan Yang
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2023, pp. 22830-22840
Recent novel view synthesis methods obtain promising results for relatively smal
l scenes, e.g., indoor environments and scenes with a few objects, but tend to f
ail for unbounded outdoor scenes with a single image as input. In this paper, we
 introduce SAMPLING, a Scene-adaptive Hierarchical Multiplane Images Representat
ion for Novel View Synthesis from a Single Image based on improved multiplane im
ages (MPI). Observing that depth distribution varies significantly for unbounded
 outdoor scenes, we employ an adaptive-bins strategy for MPI to arrange planes i
n accordance with each scene image. To represent intricate geometry and multi-sc
ale details, we further introduce a hierarchical refinement branch, which result
s in high-quality synthesized novel views. Our method demonstrates considerable
performance gains in synthesizing large-scale unbounded outdoor scenes using a s
ingle image on the KITTI dataset and generalizes well to the unseen Tanks and Te
mples dataset. The code and models will be made available at https://pkuvdig.git
hub.io/SAMPLING/.
****

DiffV2S: Diffusion-Based Video-to-Speech Synthesis with Vision-Guided Speaker Embedding
Jeongsoo Choi, Joanna Hong, Yong Man Ro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7812-7821
Recent research has demonstrated impressive results in video-to-speech synthesis which involves reconstructing speech solely from visual input. However, previous works have struggled to accurately synthesize speech due to a lack of sufficient guidance for the model to infer the correct content with the appropriate sound. To resolve the issue, they have adopted an extra speaker embedding as a speaking style guidance from a reference auditory information. Nevertheless, it is not always possible to obtain the audio information from the corresponding video input, especially during the inference time. In this paper, we present a novel vision-guided speaker embedding extractor using a self-supervised pre-trained model and prompt tuning technique. In doing so, the rich speaker embedding information can be produced solely from input visual information, and the extra audio information is not necessary during the inference time. Using the extracted vision-guided speaker embedding representations, we further develop a diffusion-based video-to-speech synthesis model, so called DiffV2S, conditioned on those speaker embeddings and the visual representation extracted from the input video. The proposed DiffV2S not only maintains phoneme details contained in the input video frames, but also creates a highly intelligible mel-spectrogram in which the speaker identities of the multiple speakers are all preserved. Our experimental results show that DiffV2S achieves the state-of-the-art performance compared to the previous video-to-speech synthesis technique.
****
PointOdyssey: A Large-Scale Synthetic Dataset for Long-Term Point Tracking
Yang Zheng, Adam W. Harley, Bokui Shen, Gordon Wetzstein, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19855-19865
We introduce PointOdyssey, a large-scale synthetic dataset, and data generation framework, for the training and evaluation of long-term fine-grained tracking algorithms. Our goal is to advance the state-of-the-art by placing emphasis on long videos with naturalistic motion. Toward the goal of naturalism, we animate deformable characters using real-world motion capture data, we build 3D scenes to match the motion capture environments, and we render camera viewpoints using trajectories mined via structure-from-motion on real videos. We create combinatorial diversity by randomizing character appearance, motion profiles, materials, lighting, 3D assets, and atmospheric effects. Our dataset currently includes 104 videos, averaging 2,000 frames long, with orders of magnitude more correspondence annotations than prior work. We show that existing methods can be trained from scratch in our dataset and outperform the published variants. Finally, we introduce modifications to the PIPs point tracking method, greatly widening its temporal receptive field, which improves its performance on PointOdyssey as well as on two real-world benchmarks. Our data and code are publicly available at: https://pointodyssey.com.
****
The Effectiveness of MAE Pre-Pretraining for Billion-Scale Pretraining
Mannat Singh, Quentin Duval, Kalyan Vasudev Alwala, Haoqi Fan, Vaibhav Aggarwal, Aaron Adcock, Armand Joulin, Piotr Dollar, Christoph Feichtenhofer, Ross Girshick, Rohit Girdhar, Ishan Misra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5484-5494
This paper revisits the standard pretrain-then-finetune paradigm used in computer vision for visual recognition tasks. Typically, state-of-the-art foundation models are pretrained using large scale (weakly) supervised datasets with billions of images. We introduce an additional pre-pretraining stage that is simple and uses the self supervised MAE technique to initialize the model. While MAE has only been shown to scale with the size of models, we find that it scales with the size of the training dataset as well. Thus, our MAE-based pre-pretraining scales with both model and data size making it applicable for training foundation models. Pre-pretraining consistently improves both the model convergence and the dow

nstream transfer performance across a range of model scales (millions to billion s of parameters), and dataset sizes (millions to billions of labels). We measure the effectiveness of pre-pretraining on 10 different visual recognition tasks spanning image classification, video recognition, object detection, low-shot classification and zero-shot recognition. Our largest model achieves new state-of-the-art results on iNaturalist-18 (91.3%), 1-shot ImageNet-1k (62.1%), and zero-shot transfer on Food-101 (96.2%). Our study reveals that model initialization plays a significant role, even for web-scale pretraining with billions of images.
****
Towards Zero Domain Gap: A Comprehensive Study of Realistic LiDAR Simulation for Autonomy Testing
Sivabalan Manivasagam, Ioan Andrei Bârsan, Jingkang Wang, Ze Yang, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8272-8282
Testing the full autonomy system in simulation is the safest and most scalable way to evaluate autonomous vehicle performance before deployment. This requires simulating sensor inputs such as LiDAR. To be effective, it is essential that the simulation has low domain gap with the real world. That is, the autonomy system in simulation should perform exactly the same way it would in the real world for the same scenario. To date, there has been limited analysis into what aspects of LiDAR phenomena affect autonomy performance. It is also difficult to evaluate the domain gap of existing LiDAR simulators, as they operate on fully synthetic scenes. In this paper, we propose a novel "paired-scenario" approach to evaluating the domain gap of a LiDAR simulator by reconstructing digital twins of real world scenarios. We can then simulate LiDAR in the scene and compare it to the real LiDAR. We leverage this setting to analyze what aspects of LiDAR simulation, such as pulse phenomena, scanning effects, and asset quality, affect the domain gap with respect to the autonomy system, including perception, prediction, and motion planning, and analyze how modifications to the simulated LiDAR influence each part. We identify key aspects that are important to model, such as motion blur, material reflectance, and the accurate geometric reconstruction of traffic participants. This helps provide research directions for improving LiDAR simulation and autonomy robustness to these effects. For more information, please visit the project website: https://waabi.ai/lidar-dg
****
GPA-3D: Geometry-aware Prototype Alignment for Unsupervised Domain Adaptive 3D Object Detection from Point Clouds
Ziyu Li, Jingming Guo, Tongtong Cao, Liu Bingbing, Wankou Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6394-6403
LiDAR-based 3D detection has made great progress in recent years. However, the performance of 3D detectors is considerably limited when deployed in unseen environments, owing to the severe domain gap problem. Existing domain adaptive 3D detection methods do not adequately consider the problem of the distributional discrepancy in feature space, thereby hindering the generalization of detectors across domains. In this work, we propose a novel unsupervised domain adaptive 3D detection framework, namely Geometry-aware Prototype Alignment (GPA-3D), which explicitly leverages the intrinsic geometric relationship from point cloud objects to reduce the feature discrepancy, thus facilitating cross-domain transferring. Specifically, GPA-3D assigns a series of tailored and learnable prototypes to point cloud objects with distinct geometric structures. Each prototype aligns BEV (bird's-eye-view) features derived from corresponding point cloud objects on source and target domains, reducing the distributional discrepancy and achieving better adaptation. The evaluation results obtained on various benchmarks, including Waymo, nuScenes and KITTI, demonstrate the superiority of our GPA-3D over the state-of-the-art approaches for different adaptation scenarios. The MindSpore version code will be publicly available at https://github.com/Liz66666/GPA3D.
****
TransHuman: A Transformer-based Human Representation for Generalizable Neural Human Rendering

Xiao Pan, Zongxin Yang, Jianxin Ma, Chang Zhou, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3544-3555

In this paper, we focus on the task of generalizable neural human rendering which trains conditional Neural Radiance Fields (NeRF) from multi-view videos of different characters. To handle the dynamic human motion, previous methods have primarily used a SparseConvNet (SPC)-based human representation to process the painted SMPL. However, such SPC-based representation i) optimizes under the volatile observation space which leads to the pose-misalignment between training and inference stages, and ii) lacks the global relationships among human parts that is critical for handling the incomplete painted SMPL. Tackling these issues, we present a brand-new framework named TransHuman, which learns the painted SMPL under the canonical space and captures the global relationships between human parts with transformers. Specifically, TransHuman is mainly composed of Transformer-based Human Encoding (TransHE), Deformable Partial Radiance Fields (DPaRF), and Fine-grained Detail Integration (FDI). TransHE first processes the painted SMPL under the canonical space via transformers for capturing the global relationships between human parts. Then, DPaRF binds each output token with a deformable radiance field for encoding the query point under the observation space. Finally, the FDI is employed to further integrate fine-grained information from reference images. Extensive experiments on ZJU-MoCap and H36M show that our TransHuman achieves a significantly new state-of-the-art performance with high efficiency. Project page: https://pansanity666.github.io/TransHuman/

****

LNPL-MIL: Learning from Noisy Pseudo Labels for Promoting Multiple Instance Learning in Whole Slide Image

Zhuchen Shao, Yifeng Wang, Yang Chen, Hao Bian, Shaohui Liu, Haoqian Wang, Yongbing Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21495-21505

Gigapixel Whole Slide Images (WSIs) aided patient diagnosis and prognosis analysis are promising directions in computational pathology. However, limited by expensive and time-consuming annotation costs, WSIs usually only have weak annotations, including 1) WSI-level Annotations (WA) and 2) Limited Patch-level Annotations (LPA). Currently, Multiple Instance Learning (MIL) often exploits WA, while LPA usually assign pseudo-labels for unlabeled data. Intuitively, pseudo-labels can serve as a practical guide for MIL, but the unreliable prediction caused by LPA inevitably introduces noise. Furthermore, WA-supervised MIL training inevitably suffers from the semantical unalignment between instances and bag-level labels. To address these problems, we design a framework called Learning from Noisy Pseudo Labels for promoting Multiple Instance Learning (LNPL-MIL), which considers both types of weak annotation. In MIL, we propose a Transformer aware of instance Order and Distribution (TOD-MIL) that strengthens instances correlation and weakens semantical unalignment in the bag. We validate our LNPL-MIL on Tumor Diagnosis and Survival Prediction, achieving state-of-the-art performance with at least 2.7%/2.9% AUC and 2.6%/2.3% C-Index improvement with the patches labeled for two scales. Ablation study and visualization analysis further verify the effectiveness.

****

Few-Shot Dataset Distillation via Translative Pre-Training

Songhua Liu, Xinchao Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18654-18664

Dataset distillation aims at a small synthetic dataset to mimic the training performance on neural networks of a given large dataset. Existing approaches heavily rely on an iterative optimization to update synthetic data and multiple forward-backward passes over thousands of neural network spaces, which introduce significant overhead for computation and are inconvenient in scenarios requiring high efficiency. In this paper, we focus on few-shot dataset distillation, where a distilled dataset is synthesized with only a few or even a single network. To this end, we introduce the notion of distillation space, such that synthetic data optimized only in this specific space can achieve the effect of those optimized through numerous neural networks, with dramatically accelerated training and redu

ced computational cost. To learn such a distillation space, we first formulate the problem as a quad-level optimization framework and propose a bi-level algorithm. Nevertheless, the algorithm in its original form has a large memory footprint in practice due to the back-propagation through an unrolled computational graph. We then convert the problem of learning the distillation space to a first-order one based on image translation. Specifically, the synthetic images are optimized in an arbitrary but fixed neural space and then translated to those in the targeted distillation space. We pre-train the translator on some large datasets like ImageNet so that it requires only a limited number of adaptation steps on the target dataset. Extensive experiments demonstrate that the translator after pre-training and a limited number of adaptation steps achieves comparable distillation performance with state of the arts, with 15x acceleration. It also exerts satisfactory generalization performance across different datasets, storage budgets, and numbers of classes.
****
Random Sub-Samples Generation for Self-Supervised Real Image Denoising
Yizhong Pan, Xiao Liu, Xiangyu Liao, Yuanzhouhan Cao, Chao Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12150-12159
With sufficient paired training samples, the supervised deep learning methods have attracted much attention in image denoising because of their superior performance. However, it is still very challenging to widely utilize the supervised methods in real cases due to the lack of paired noisy-clean images. Meanwhile, most self-supervised denoising methods are ineffective as well when applied to the real-world denoising tasks because of their strict assumptions in applications. For example, as a typical method for self-supervised denoising, the original blind spot network (BSN) assumes that the noise is pixel-wise independent, which is much different from the real cases. To solve this problem, we propose a novel self-supervised real image denoising framework named Sampling Difference As Perturbation (SDAP) based on Random Sub-samples Generation (RSG) with a cyclic sample difference loss. Specifically, we dig deeper into the properties of BSN to make it more suitable for real noise. Surprisingly, we find that adding an appropriate perturbation to the training images can effectively improve the performance of BSN. Further, we propose that the sampling difference can be considered as perturbation to achieve better results. Finally we propose a new BSN framework in combination with our RSG strategy. The results show that it significantly outperforms other state-of-the-art self-supervised denoising methods on real-world datasets. The code is available at https://github.com/p1y2z3/SDAP.
****
Waffling Around for Performance: Visual Classification with Random Words and Broad Concepts
Karsten Roth, Jae Myung Kim, A. Sophia Koepke, Oriol Vinyals, Cordelia Schmid, Zeynep Akata; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15746-15757
The visual classification performance of vision-language models such as CLIP has been shown to benefit from additional semantic knowledge from large language models (LLMs) such as GPT-3. In particular, averaging over LLM-generated class descriptors, e.g. "waffle, which has a round shape", can notably improve generalization performance. In this work, we critically study this behavior and propose WaffleCLIP, a framework for zero-shot visual classification which simply replaces LLM-generated descriptors with random character and word descriptors. Without querying external models, we achieve comparable performance gains on a large number of visual classification tasks. This allows WaffleCLIP to both serve as a low-cost alternative, as well as a sanity check for any future LLM-based vision-language model extensions. We conduct an extensive experimental study on the impact and shortcomings of additional semantics introduced with LLM-generated descriptors, and showcase how - if available - semantic context is better leveraged by querying LLMs for high-level concepts, which we show can be done to jointly resolve potential class name ambiguities. Code is available here: https://github.com/Expla

inableML/WaffleCLIP.
****
Unsupervised Surface Anomaly Detection with Diffusion Probabilistic Model
Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, Shu-Tao Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6782-6791
Unsupervised surface anomaly detection aims at discovering and localizing anomalous patterns using only anomaly-free training samples. Reconstruction-based models are among the most popular and successful methods, which rely on the assumption that anomaly regions are more difficult to reconstruct. However, there are three major challenges to the practical application of this approach: 1) the reconstruction quality needs to be further improved since it has a great impact on the final result, especially for images with structural changes; 2) it is observed that for many neural networks, the anomalies can also be well reconstructed, which severely violates the underlying assumption; 3) since reconstruction is an ill-conditioned problem, a test instance may correspond to multiple normal patterns, but most current reconstruction-based methods have ignored this critical fact. In this paper, we propose DiffAD, a method for unsupervised anomaly detection based on the latent diffusion model, inspired by its ability to generate high-quality and diverse images. We further propose noisy condition embedding and interpolated channels to address the aforementioned challenges in the general reconstruction-based pipeline. Extensive experiments show that our method achieves state-of-the-art performance on the challenging MVTec dataset, especially in localization accuracy.
****
AutoAD II: The Sequel - Who, When, and What in Movie Audio Description
Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13645-13655
Audio Description (AD) is the task of generating descriptions of visual content, at suitable time intervals, for the benefit of visually impaired audiences. For movies, this presents notable challenges -- AD must occur only during existing pauses in dialogue, should refer to characters by name, and ought to aid understanding of the storyline as a whole.
To this end, we develop a new model for automatically generating movie AD, given CLIP visual features of the frames, the cast list, and the temporal locations of the speech; addressing all three of the `who', `when', and `what' questions: (i) who -- we introduce a character bank consisting of the character's name, the actor that played the part, and a CLIP feature of their face, for the principal cast of each movie, and demonstrate how this can be used to improve naming in the generated AD; (ii) when -- we investigate several models for determining whether an AD should be generated for a time interval or not, based on the visual content of the interval and its neighbours; and (iii) what -- we implement a new vision-language model for this task, that can ingest the proposals from the character bank, whilst conditioning on the visual features using cross-attention, and demonstrate how this improves over previous architectures for AD text generation in an apples-to-apples comparison.
****
TinyCLIP: CLIP Distillation via Affinity Mimicking and Weight Inheritance
Kan Wu, Houwen Peng, Zhenghong Zhou, Bin Xiao, Mengchen Liu, Lu Yuan, Hong Xuan, Michael Valenzuela, Xi (Stephen) Chen, Xinggang Wang, Hongyang Chao, Han Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21970-21980
In this paper, we propose a novel cross-modal distillation method, called TinyCLIP, for large-scale language-image pre-trained models. The method introduces two core techniques: affinity mimicking and weight inheritance. Affinity mimicking explores the interaction between modalities during distillation, enabling student models to mimic teachers' behavior of learning cross-modal feature alignment in a visual-linguistic affinity space. Weight inheritance transmits the pre-trained weights from the teacher models to their student counterparts to improve dist

illation efficiency. Moreover, we extend the method into a multi-stage progressive distillation to mitigate the loss of informative weights during extreme compression. Comprehensive experiments demonstrate the efficacy of TinyCLIP, showing that it can reduce the size of the pre-trained CLIP ViT-B/32 by 50%, while maintaining comparable zero-shot performance. While aiming for comparable performance, distillation with weight inheritance can speed up the training by 1.4 - 7.8x compared to training from scratch. Moreover, our TinyCLIP ViT-8M/16, trained on YFCC-15M, achieves an impressive zero-shot top-1 accuracy of 41.1% on ImageNet, surpassing the original CLIP ViT-B/16 by 3.5% while utilizing only 8.9% parameters. Finally, we demonstrate the good transferability of TinyCLIP in various downstream tasks. Code and models will be open-sourced at aka.ms/tinyclip.
****
Hyperbolic Chamfer Distance for Point Cloud Completion
Fangzhou Lin, Yun Yue, Songlin Hou, Xuechu Yu, Yajun Xu, Kazunori D Yamada, Ziming Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14595-14606
Chamfer distance (CD) is a standard metric to measure the shape dissimilarity between point clouds in point cloud completion, as well as a loss function for (deep) learning. However, it is well known that CD is vulnerable to outliers, leading to the drift towards suboptimal models. In contrast to the literature where most works address such issues in Euclidean space, we propose an extremely simple yet powerful metric for point cloud completion, namely Hyperbolic Chamfer Distance (HyperCD), that computes CD in hyperbolic space. In backpropagation, HyperCD consistently assigns higher weights to the matched point pairs with smaller Euclidean distances. In this way, good point matches are likely to be preserved while bad matches can be updated gradually, leading to better completion results. We demonstrate state-of-the-art performance on the benchmark datasets, i.e. PCN, ShapeNet-55, and ShapeNet-34, and show from visualization that HyperCD can significantly improve the surface smoothness. Code is available at: https://github.com/Zhang-VISLab.
****
Democratising 2D Sketch to 3D Shape Retrieval Through Pivoting
Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, Yi-Zhe Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23275-23286
This paper studies the problem of 2D sketch to 3D shape retrieval, but with a focus on democratising the process. We would like this democratisation to happen on two fronts: (i) to remove the need for large-scale specifically sourced 2D sketch and 3D shape datasets, and (ii) to remove restrictions on how well the user needs to sketch and from what viewpoint. The end result is a system that is trainable using existing datasets, and once trained allows users to sketch regardless of drawing skills and without restriction on view angle. We achieve all this via a clever use of pivoting, along with novel designs that injects 3D understanding of 2D sketches into the system. We perform pivoting on two existing datasets, each from a distant research domain to the other: 2D sketch and photo pairs from the sketch-based image retrieval field (SBIR), and 3D shapes from ShapeNet. It follows that the actual feature pivoting happens on photos from the former and 2D projections from the latter. Doing this already achieves most of our democratisation challenge -- the level of 2D sketch abstraction embedded in SBIR dataset offers demoralization on drawing quality, and the whole thing works without a specifically sourced 2D sketch and 3D model pair. To further achieve democratisation on sketching viewpoint, we "lift" 2D sketches to 3D space using Blind Perspective-n-Points (BPnP) that injects 3D-aware information into the sketch encoder. Results show ours achieves competitive performance compared with fully-supervised baselines, while meeting all set democratisation goals.
****
Simoun: Synergizing Interactive Motion-appearance Understanding for Vision-based Reinforcement Learning
Yangru Huang, Peixi Peng, Yifan Zhao, Yunpeng Zhai, Haoran Xu, Yonghong Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20

23, pp. 176-185
Efficient motion and appearance modeling are critical for vision-based Reinforcement Learning (RL). However, existing methods struggle to reconcile motion and appearance information within the state representations learned from a single observation encoder. To address the problem, we present Synergizing Interactive Motion-appearance Understanding (Simoun), a unified framework for vision-based RL. Given consecutive observation frames, Simoun deliberately and interactively learns both motion and appearance features through a dual-path network architecture. The learning process collaborates with a structural interactive module, which explores the latent motion-appearance structures from the two network paths to leverage their complementarity. To promote sample efficiency, we further design a consistency-guided curiosity module to encourage the exploration of under-learned observations. During training, the curiosity module provides intrinsic rewards according to the consistency of environmental temporal dynamics, which are deduced from both motion and appearance network paths. Experiments conducted on the DeepMind control suite and CARLA automatic driving benchmarks demonstrate the effectiveness of Simoun, where it performs favorably against state-of-the-art methods.
****
AG3D: Learning to Generate 3D Avatars from 2D Image Collections
Zijian Dong, Xu Chen, Jinlong Yang, Michael J. Black, Otmar Hilliges, Andreas Geiger; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14916-14927
While progress in 2D generative models of human appearance has been rapid, many applications require 3D avatars that can be animated and rendered. Unfortunately, most existing methods for learning generative models of 3D humans with diverse shape and appearance require 3D training data, which is limited and expensive to acquire. The key to progress is hence to learn generative models of 3D avatars from abundant unstructured 2D image collections. However, learning realistic and complete 3D appearance and geometry in this under-constrained setting remains challenging, especially in the presence of loose clothing such as dresses. In this paper, we propose a new adversarial generative model of realistic 3D people from 2D images. Our method captures shape and deformation of the body and loose clothing by adopting a holistic 3D generator and integrating an efficient, flexible, articulation module. To improve realism, we train our model using multiple discriminators while also integrating geometric cues in the form of predicted 2D normal maps. We experimentally find that our method outperforms previous 3D- and articulation-aware methods in terms of geometry and appearance. We validate the effectiveness of our model and the importance of each component via systematic ablation studies.
****
KECOR: Kernel Coding Rate Maximization for Active 3D Object Detection
Yadan Luo, Zhuoxiao Chen, Zhen Fang, Zheng Zhang, Mahsa Baktashmotlagh, Zi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18279-18290
Achieving a reliable LiDAR-based object detector in autonomous driving is paramount, but its success hinges on obtaining large amounts of precise 3D annotations. Active learning (AL) seeks to mitigate the annotation burden through algorithms that use fewer labels and can attain performance comparable to fully supervised learning. Although AL has shown promise, current approaches prioritize the selection of unlabeled point clouds with high aleatoric and/or epistemic uncertainty, leading to the selection of more instances for labeling and reduced computational efficiency. In this paper, we resort to a novel kernel coding rate maximization (KECOR) strategy which aims to identify the most informative point clouds to acquire labels through the lens of information theory. Greedy search is applied to seek desired point clouds that can maximize the minimal number of bits required to encode the latent features. To determine the uniqueness and informativeness of the selected samples from the model perspective, we construct a proxy network of the 3D detector head and compute the outer product of Jacobians from all proxy layers to form the empirical neural tangent kernel (NTK) matrix. To accom

modate both one-stage (i.e., SECOND) and two-stage detectors (i.e., PV-RCNN), we further incorporate the classification entropy maximization and well trade-off between detection performance and the total number of bounding boxes selected for annotation. Extensive experiments conducted on two 3D benchmarks and a 2D detection dataset evidence the superiority and versatility of the proposed approach. Our results show that approximately 44% box-level annotation costs and 26% computational time are reduced compared to the state-of-the-art AL method, without compromising detection performance.
****
Learned Image Reasoning Prior Penetrates Deep Unfolding Network for Panchromatic and Multi-spectral Image Fusion
Man Zhou, Jie Huang, Naishan Zheng, Chongyi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12398-12407
The success of deep neural networks for pan-sharpening is commonly in a form of black box, lacking transparency and interpretability. To alleviate this issue, we propose a novel model-driven deep unfolding framework with image reasoning prior tailored for the pan-sharpening task. Different from existing unfolding solutions that deliver the proximal operator networks as the uncertain and vague priors, our framework is motivated by the content reasoning ability of masked autoencoders (MAE) with insightful designs. Specifically, the pre-trained MAE with spatial masking strategy, acting as intrinsic reasoning prior, is embedded into unfolding architecture. Meanwhile, the pre-trained MAE with spatial-spectral masking strategy is treated as the regularization term within loss function to constrain the spatial-spectral consistency. Such designs penetrate the image reasoning prior into deep unfolding networks while improving its interpretability and representation capability. The uniqueness of our framework is that the holistic learning process is explicitly integrated with the inherent physical mechanism underlying the pan-sharpening task. Extensive experiments on multiple satellite datasets demonstrate the superiority of our method over the existing state-of-the-art approaches.
****
Representation Disparity-aware Distillation for 3D Object Detection
Yanjing Li, Sheng Xu, Mingbao Lin, Jihao Yin, Baochang Zhang, Xianbin Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6715-6724
In this paper, we focus on developing knowledge distillation (KD) for compact 3D detectors. We observe that off-the-shelf KD methods manifest their efficacy only when the teacher model and student counterpart share similar intermediate feature representations. This might explain why they are less effective in building extreme-compact 3D detectors where significant representation disparity arises due primarily to the intrinsic sparsity and irregularity in 3D point clouds. This paper presents a novel representation disparity-aware distillation (RDD) method to address the representation disparity issue and reduce performance gap between compact students and over-parameterized teachers. This is accomplished by building our RDD from an innovative perspective of information bottleneck (IB), which can effectively minimize the disparity of proposal region pairs from student and teacher in features and logits. Extensive experiments are performed to demonstrate the superiority of our RDD over existing KD methods. For example, our RDD increases mAP of CP-Voxel-S to 57.1% on nuScenes dataset, which even surpasses teacher performance while taking up only 42% FLOPs.
****
NCHO: Unsupervised Learning for Neural 3D Composition of Humans and Objects
Taeksoo Kim, Shunsuke Saito, Hanbyul Joo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14817-14828
Deep generative models have been recently extended to synthesizing 3D digital humans. However, previous approaches treat clothed humans as a single chunk of geometry without considering the compositionality of clothing and accessories. As a result, individual items cannot be naturally composed into novel identities, leading to limited expressiveness and controllability of generative 3D avatars. While several methods attempt to address this by leveraging synthetic data, the in

teraction between humans and objects is not authentic due to the domain gap, and manual asset creation is difficult to scale for a wide variety of objects. In this work, we present a novel framework for learning a compositional generative model of humans and objects (backpacks, coats, scarves, and more) from real-world 3D scans. Our compositional model is interaction-aware, meaning the spatial relationship between humans and objects, and the mutual shape change by physical contact is fully incorporated. The key challenge is that, since humans and objects are in contact, their 3D scans are merged into a single piece. To decompose them without manual annotations, we propose to leverage two sets of 3D scans of a single person with and without objects. Our approach learns to decompose objects and naturally compose them back into a generative human model in an unsupervised manner. Despite our simple setup requiring only the capture of a single subject with objects, our experiments demonstrate the strong generalization of our model by enabling the natural composition of objects to diverse identities in various poses and the composition of multiple objects, which is unseen in training data. The project page is available at https://taeksuu.github.io/ncho.
****
Breaking The Limits of Text-conditioned 3D Motion Synthesis with Elaborative Descriptions
Yijun Qian, Jack Urbanek, Alexander G. Hauptmann, Jungdam Won; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2306-2316
Given its wide applications, there is increasing focus on generating 3D human motions from textual descriptions. Differing from the majority of previous works, which regard actions as single entities and can only generate short sequences for simple motions, we propose EMS, an elaborative motion synthesis model conditioned on detailed natural language descriptions. It generates natural and smooth motion sequences for long and complicated actions by factorizing them into groups of atomic actions. Meanwhile, it understands atomic-action level attributes (e.g., motion direction, speed, and body parts) and enables users to generate sequences of unseen complex actions from unique sequences of known atomic actions with independent attribute settings and timings applied. We evaluate our method on the KIT Motion-Language and BABEL benchmarks, where it outperforms all previous state-of-the-art with noticeable margins.
****
VL-PET: Vision-and-Language Parameter-Efficient Tuning via Granularity Control
Zi-Yuan Hu, Yanyang Li, Michael R. Lyu, Liwei Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3010-3020
As the model size of pre-trained language models (PLMs) grows rapidly, full fine-tuning becomes prohibitively expensive for model training and storage. In vision-and-language (VL), parameter-efficient tuning (PET) techniques are proposed to integrate modular modifications (e.g., Adapter) into encoder-decoder PLMs. By tuning a small set of trainable parameters, these techniques perform on par with full fine-tuning. However, excessive modular modifications and neglecting the unique abilities of the encoders and decoders can lead to performance degradation, while existing PET techniques (e.g., VL-Adapter) overlook these issues. In this paper, we propose a Vision-and-Language Parameter-Efficient Tuning (VL-PET) framework to impose effective control over modular modifications via a novel granularity-controlled mechanism. Considering different granularity-controlled matrices generated by this mechanism, a variety of model-agnostic VL-PET modules can be instantiated from our framework for better efficiency and effectiveness trade-offs. We further propose lightweight designs to enhance VL alignment and modeling for the encoders and maintain text generation for the decoders. Extensive experiments conducted on four image-text tasks and four video-text tasks demonstrate the efficiency, effectiveness, scalability and transferability of our VL-PET framework. In particular, our VL-PET-large significantly outperforms full fine-tuning by 2.39% (2.61%) and VL-Adapter by 2.92% (3.41%) with BART-base (T5-base) on image-text tasks, while utilizing fewer trainable parameters. Furthermore, we validate the enhanced effect of employing our VL-PET designs (e.g., granularity-controlled mechanism and lightweight designs) on existing PET techniques, enabling

them to achieve significant performance improvements.
****
ROME: Robustifying Memory-Efficient NAS via Topology Disentanglement and Gradient Accumulation
Xiaoxing Wang, Xiangxiang Chu, Yuda Fan, Zhexi Zhang, Bo Zhang, Xiaokang Yang, Junchi Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5939-5949
Albeit being a prevalent architecture searching approach, differentiable architecture search (DARTS) is largely hindered by its substantial memory cost since the entire supernet resides in the memory. This is where the single-path DARTS comes in, which only chooses a single-path submodel at each step. While being memory-friendly, it also comes with low computational costs. Nonetheless, we discover a critical issue of single-path DARTS that has not been primarily noticed. Namely, it also suffers from severe performance collapse since too many parameter-free operations like skip connections are derived, just like DARTS does. In this paper, we propose a new algorithm called RObustifying Memory-Efficient NAS (ROME) to give a cure. First, we disentangle the topology search from the operation search to make searching and evaluation consistent. We then adopt Gumbel-Top2 reparameterization and gradient accumulation to robustify the unwieldy bi-level optimization. We verify ROME extensively across 15 benchmarks to demonstrate its effectiveness and robustness.
****
Toward Multi-Granularity Decision-Making: Explicit Visual Reasoning with Hierarchical Knowledge
Yifeng Zhang, Shi Chen, Qi Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2573-2583
Answering visual questions requires the ability to parse visual observations and correlate them with a variety of knowledge. Existing visual question answering (VQA) models either pay little attention to the role of knowledge or do not take into account the granularity of knowledge, e.g., attaching the color of "grassland" to "ground"). They have yet to develop the capability of modeling knowledge of multiple granularity, and are also vulnerable to spurious data biases. To fill the gap, this paper makes progresses from two distinct perspectives: (1) It presents a Hierarchical Concept Graph (HCG) that discriminates and associates multi-granularity concepts with a multi-layered hierarchical structure, aligning visual observations with knowledge across different levels to alleviate data biases. (2) To facilitate a comprehensive understanding of how knowledge contributes throughout the decision-making process, we further propose an interpretable Hierarchical Concept Neural Module Network (HCNMN) It explicitly propagates multi-granularity knowledge across the hierarchical structure and incorporates them with a sequence of reasoning steps, providing a transparent interface to elaborate on the integration of observations and knowledge. Through extensive experiments on multiple challenging datasets (i.e., GQA,VQA,FVQA,OK-VQA) , we demonstrate the effectiveness of our method in answering questions in different scenarios. Our code is available at https://github.com/SuperJohnZhang/HCNMN.
****
3D-aware Image Generation using 2D Diffusion Models
Jianfeng Xiang, Jiaolong Yang, Binbin Huang, Xin Tong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2383-2393
In this paper, we introduce a novel 3D-aware image generation method that leverages 2D diffusion models. We formulate the 3D-aware image generation task as multiview 2D image set generation, and further to a sequential unconditional-conditional multiview image generation process. This allows us to utilize 2D diffusion models to boost the generative modelling power of the method. Additionally, we incorporate depth information from monocular depth estimators to construct the training data for the conditional diffusion model using only still images. We train our method on a large-scale unstructured dataset, i.e., ImageNet, which is not addressed by previous methods. It produces high-quality images that significantly outperform prior methods. Furthermore, our approach showcases its capability to generate instances with large view angles, even though the training images ar

e diverse and unaligned, gathered from "in-the-wild" real-world environments.
****
Locating Noise is Halfway Denoising for Semi-Supervised Segmentation
Yan Fang, Feng Zhu, Bowen Cheng, Luoqi Liu, Yao Zhao, Yunchao Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16612-16622
We investigate semi-supervised semantic segmentation with self-training, where a teacher model generates pseudo masks to exploit the benefits of a large amount of unlabeled images. We notice that the noisy label from the generated pseudo masks is the major obstacle to achieving good performance. Previous works all treat the noise in pixel level and ignore the contextual information of the noise. This work shows that locating the patch-wise noisy region is a better way to deal with noise. To be specific, our method, named Uncertainty-aware Patch CutMix (UPC), first estimates the uncertainty of per-pixel prediction for pseudo masks of unlabeled images. Then UPC splits the uncertainty map into patches and calculates patch-wise uncertainty. UPC selects top-k most uncertain patches to generate the uncertain regions. Finally, uncertain regions are replaced with reliable ones from labeled images. We conduct extensive experiments using standard semi-supervised settings on Pascal VOC and Cityscapes. Experiment results show that UPC can significantly boost the performance of the state-of-the-art methods. In addition, we further demonstrate that our UPC is robust to out-of-distribution unlabeled images, eg, MSCOCO.
****
Learning Non-Local Spatial-Angular Correlation for Light Field Image Super-Resolution
Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Shilin Zhou, Yulan Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12376-12386
Exploiting spatial-angular correlation is crucial to light field (LF) image super-resolution (SR), but is highly challenging due to its non-local property caused by the disparities among LF images. Although many deep neural networks (DNNs) have been developed for LF image SR and achieved continuously improved performance, existing methods cannot well leverage the long-range spatial-angular correlation and thus suffer a significant performance drop when handling scenes with large disparity variations. In this paper, we propose a simple yet effective method to learn the non-local spatial-angular correlation for LF image SR. In our method, we adopt the epipolar plane image (EPI) representation to project the 4D spatial-angular correlation onto multiple 2D EPI planes, and then develop a Transformer network with repetitive self-attention operations to learn the spatial-angular correlation by modeling the dependencies between each pair of EPI pixels. Our method can fully incorporate the information from all angular views while achieving a global receptive field along the epipolar line. We conduct extensive experiments with insightful visualizations to validate the effectiveness of our method. Comparative results on five public datasets show that our method not only achieves state-of-the-art SR performance, but also performs robust to disparity variations.
****
ICE-NeRF: Interactive Color Editing of NeRFs via Decomposition-Aware Weight Optimization
Jae-Hyeok Lee, Dae-Shik Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3491-3501
Neural Radiance Fields (NeRFs) have gained considerable attention for their high-quality results in 3D scene reconstruction and rendering. Recently, there have been active studies on various tasks such as novel view synthesis and scene editing. However, editing NeRFs is challenging as accurately decomposing the desired area of 3D space and ensuring the consistency of edited results from different angles is difficult. In this paper, we propose ICE-NeRF, an Interactive Color Editing framework that performs color editing by taking a pre-trained NeRF and a rough user mask as input. Our proposed method performs the entire color editing process in only under a minute using a partial fine-tuning approach. To perform e

ffective color editing, we address two issues: (1) the entanglement of the implicit representation that causes unwanted color changes in undesired areas when learning weights, and (2) the loss of multi-view consistency when fine-tuning for a single or a few views. To address these issues, we introduce two techniques: Activation Field-based Regularization (AFR) and Single-mask Multi-view Rendering (SMR). The AFR performs weight regularization during fine-tuning based on the assumption that not all weights have an equal impact on the desired area. The SMR maps the 2D mask to 3D space through inverse projection and renders it from other views to generate multi-view masks. ICE-NeRF not only enables well-decomposed, multi-view consistent color editing but also significantly reduces processing time compared to existing methods.
****

SPANet: Frequency-balancing Token Mixer using Spectral Pooling Aggregation Modulation

Guhnoo Yun, Juhan Yoo, Kijung Kim, Jeongho Lee, Dong Hwan Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6113-6124

Recent studies show that self-attentions behave like low-pass filters (as opposed to convolutions) and enhancing their high-pass filtering capability improves model performance. Contrary to this idea, we investigate existing convolution-based models with spectral analysis and observe that improving the low-pass filtering in convolution operations also leads to performance improvement. To account for this observation, we hypothesize that utilizing optimal token mixers that capture balanced representations of both high- and low-frequency components can enhance the performance of models. We verify this by decomposing visual features into the frequency domain and combining them in a balanced manner. To handle this, we replace the balancing problem with a mask filtering problem in the frequency domain. Then, we introduce a novel token-mixer named SPAM and leverage it to derive a MetaFormer model termed as SPANet. Experimental results show that the proposed method provides a way to achieve this balance, and
the balanced representations of both high- and low-frequency components can improve the performance of models on multiple computer vision tasks. Our code is available at https://doranlyong.github.io/projects/spanet/.
****

ASAG: Building Strong One-Decoder-Layer Sparse Detectors via Adaptive Sparse Anchor Generation

Shenghao Fu, Junkai Yan, Yipeng Gao, Xiaohua Xie, Wei-Shi Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6328-6338

Recent sparse detectors with multiple, e.g. six, decoder layers achieve promising performance but much inference time due to complex heads. Previous works have explored using dense priors as initialization and built one-decoder-layer detectors. Although they gain remarkable acceleration, their performance still lags behind their six-decoder-layer counterparts by a large margin. In this work, we aim to bridge this performance gap while retaining fast speed. We find that the architecture discrepancy between dense and sparse detectors leads to feature conflict, hampering the performance of one-decoder-layer detectors. Thus we propose Adaptive Sparse Anchor Generator (ASAG) which predicts dynamic anchors on patches rather than grids in a sparse way so that it alleviates the feature conflict problem. For each image, ASAG dynamically selects which feature maps and which locations to predict, forming a fully adaptive way to generate image-specific anchors. Further, a simple and effective Query Weighting method eases the training instability from adaptiveness. Extensive experiments show that our method outperforms dense-initialized ones and achieves a better speed-accuracy trade-off. The code is available at https://github.com/iSEE-Laboratory/ASAG.
****

MGMAE: Motion Guided Masking for Video Masked Autoencoding

Bingkun Huang, Zhiyu Zhao, Guozhen Zhang, Yu Qiao, Limin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13493-13504

Masked autoencoding has shown excellent performance on self-supervised video representation learning. Temporal redundancy has led to a high masking ratio and customized masking strategy in VideoMAE. In this paper, we aim to further improve the performance of video masked autoencoding by introducing a motion guided masking strategy. Our key insight is that motion is a general and unique prior in video, which should be taken into account during masked pre-training. Our motion guided masking explicitly incorporates motion information to build temporal consistent masking volume. Based on this masking volume, we can track the unmasked tokens in time and sample a set of temporal consistent cubes from videos. These temporal aligned unmasked tokens will further relieve the information leakage issue in time and encourage the MGMAE to learn more useful structure information. We implement our MGMAE with an online efficient optical flow estimator and backward masking map warping strategy. We perform experiments on the datasets of Something-Something V2 and Kinetics-400, demonstrating the superior performance of our MGMAE to the original VideoMAE. In addition, we provide the visualization analysis to illustrate that our MGMAE can sample temporal consistent cubes in a motion-adaptive manner for more effective video pre-training.
****
The Perils of Learning From Unlabeled Data: Backdoor Attacks on Semi-supervised Learning
Virat Shejwalkar, Lingjuan Lyu, Amir Houmansadr; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4730-4740
Semi-supervised learning (SSL) is gaining popularity as it reduces cost of machine learning (ML) by training high performance models using unlabeled data. In this paper, we reveal that the key feature of SSL, i.e., learning from (non-inspected) unlabeled data, exposes SSL to strong poisoning attacks that can significantly damage its security. Poisoning is a long-standing problem in conventional supervised ML, but we argue that, as SSL relies on non-inspected unlabeled data, poisoning poses a more significant threat to SSL. We demonstrate this by designing a backdoor poisoning attack on SSL that can be conducted by a weak adversary with no knowledge of the target SSL pipeline. This is unlike prior poisoning attacks on supervised ML that assume strong adversaries with impractical capabilities. We show that by poisoning only 0.2% of the unlabeled training data, our (weak) adversary can successfully cause misclassification on more than 80% of test inputs (when they contain the backdoor trigger). Our attack remains effective across different benchmark datasets and SSL algorithms, and even circumvents state-of-the-art defenses against backdoor attacks. Our work raises significant concerns about the security of SSL in real-world security critical applications.
****
SSB: Simple but Strong Baseline for Boosting Performance of Open-Set Semi-Supervised Learning
Yue Fan, Anna Kukleva, Dengxin Dai, Bernt Schiele; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16068-16078
Semi-supervised learning (SSL) methods effectively leverage unlabeled data to improve model generalization. However, SSL models often underperform in open-set scenarios, where unlabeled data contain outliers from novel categories that do not appear in the labeled set. In this paper, we study the challenging and realistic open-set SSL setting, where the goal is to both correctly classify inliers and to detect outliers. Intuitively, the inlier classifier should be trained on inlier data only. However, we find that inlier classification performance can be largely improved by incorporating high-confidence pseudo-labeled data, regardless of whether they are inliers or outliers. Also, we propose to utilize non-linear transformations to separate the features used for inlier classification and outlier detection in the multi-task learning framework, preventing adverse effects between them. Additionally, we introduce pseudo-negative mining, which further boosts outlier detection performance. The three ingredients lead to what we call Simple but Strong Baseline (SSB) for open-set SSL. In experiments, SSB greatly improves both inlier classification and outlier detection performance, outperforming existing methods by a large margin. Our code will be released at https://github.com/YUE-FAN/SSB.

****
StyleDiffusion: Controllable Disentangled Style Transfer via Diffusion Models
Zhizhong Wang, Lei Zhao, Wei Xing; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7677-7689
Content and style (C-S) disentanglement is a fundamental problem and critical challenge of style transfer. Existing approaches based on explicit definitions (e.g., Gram matrix) or implicit learning (e.g., GANs) are neither interpretable nor easy to control, resulting in entangled representations and less satisfying results. In this paper, we propose a new C-S disentangled framework for style transfer without using previous assumptions. The key insight is to explicitly extract the content information and implicitly learn the complementary style information, yielding interpretable and controllable C-S disentanglement and style transfer. A simple yet effective CLIP-based style disentanglement loss coordinated with a style reconstruction prior is introduced to disentangle C-S in the CLIP image space. By further leveraging the powerful style removal and generative ability of diffusion models, our framework achieves superior results than state of the art and flexible C-S disentanglement and trade-off control. Our work provides new insights into the C-S disentanglement in style transfer and demonstrates the potential of diffusion models for learning well-disentangled C-S characteristics.
****
AdvDiffuser: Natural Adversarial Example Synthesis with Diffusion Models
Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, Cheng-Zhong Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4562-4572
Previous work on adversarial examples typically involves a fixed norm perturbation budget, which fails to capture the way humans perceive perturbations. Recent work has shifted towards investigating natural unrestricted adversarial examples (UAEs) that breaks l_p perturbation bounds but nonetheless remain semantically plausible. Current methods use GAN or VAE to generate UAEs by perturbing latent codes. However, this leads to loss of high-level information, resulting in low-quality and unnatural UAEs. In light of this, we propose AddDiffuser, a new method for synthesizing natural UAEs using diffusion models. It can generate UAEs from scratch or conditionally based on reference images. To generate natural UAEs, we perturb predicted images to steer their latent code towards the adversarial sample space of a particular classifier. In addition, we propose adversarial inpainting based on class activation mapping to retain the salient regions of the image while perturbing less important areas. Our method achieves impressive results on CIFAR-10, CelebA and ImageNet, and we demonstrate that it can defeat the most robust models on the RobustBench leaderboard with near 100% success rates. Furthermore, The synthesized UAEs are not only more natural but also stronger compared to the current state-of-the-art attacks. Specifically, compared with GA-attack, the UAEs generated with AdvDiffuser exhibit 6xsmaller LPIPS perturbations, 2 ~ 3 xsmaller FID scores and 0.28 higher in SSIM metrics, making them perceptually stealthier. Lastly, it is capable of generating an unlimited number of natural adversarial examples. For more please visit our project page: Link to follow.
****
ViewRefer: Grasp the Multi-view Knowledge for 3D Visual Grounding
Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, Xuelong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15372-15383
Understanding 3D scenes from multi-view inputs has been proven to alleviate the view discrepancy issue in 3D visual grounding. However, existing methods normally neglect the view cues embedded in the text modality and fail to weigh the relative importance of different views. In this paper, we propose ViewRefer, a multi-view framework for 3D visual grounding exploring how to grasp the view knowledge from both text and 3D modalities. For the text branch, ViewRefer leverages the diverse linguistic knowledge of large-scale language models to expand a single grounding text to multiple geometry-consistent descriptions. Meanwhile, in the 3D modality, a transformer fusion module with inter-view attention is introduced to boost the interaction of objects across views. On top of that, we further pre

sent a set of learnable multi-view prototypes, which memorize scene-agnostic knowledge for different views, and enhance the framework from two perspectives: a view-guided attention module for more robust text features, and a view-guided scoring strategy during the final prediction. With our designed paradigm, ViewRefer achieves superior performance on three benchmarks and surpasses the second-best by +2.8%, +1.5%, and +1.35% on Sr3D, Nr3D, and ScanRefer.
****
CaPhy: Capturing Physical Properties for Animatable Human Avatars
Zhaoqi Su, Liangxiao Hu, Siyou Lin, Hongwen Zhang, Shengping Zhang, Justus Thies, Yebin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14150-14160
We present CaPhy, a novel method for reconstructing animatable human avatars with realistic dynamic properties for clothing. Specifically, we aim for capturing the geometric and physical properties of the clothing from real observations. This allows us to apply novel poses to the human avatar with physically correct deformations and wrinkles of the clothing. To this end, we combine unsupervised training with physics-based losses and 3D-supervised training using scanned data to reconstruct a dynamic model of clothing that is physically realistic and conforms to the human scans. We also optimize the physical parameters of the underlying physical model from the scans by introducing gradient constraints of the physics-based losses. In contrast to previous work on 3D avatar reconstruction, our method is able to generalize to novel poses with realistic dynamic cloth deformations. Experiments on several subjects demonstrate that our method can estimate the physical properties of the garments, resulting in superior quantitative and qualitative results compared with previous methods.
****
DarSwin: Distortion Aware Radial Swin Transformer
Akshaya Athwale, Arman Afrasiyabi, Justin Lagüe, Ichrak Shili, Ola Ahmad, Jean-François Lalonde; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5929-5938
Wide-angle lenses are commonly used in perception tasks requiring a large field of view. Unfortunately, these lenses produce significant distortions making conventional models that ignore the distortion effects unable to adapt to wide-angle images. In this paper, we present a novel transformer-based model that automatically adapts to the distortion produced by wide-angle lenses. We leverage the physical characteristics of such lenses, which are analytically defined by the radial distortion profile (assumed to be known), to develop a distortion aware radial swin transformer (DarSwin). In contrast to conventional transformer-based architectures, DarSwin comprises a radial patch partitioning, a distortion-based sampling technique for creating token embeddings, and an angular position encoding for radial patch merging. We validate our method on classification tasks using synthetically distorted ImageNet data and show through extensive experiments that DarSwin can perform zero-shot adaptation to unseen distortions of different wide-angle lenses. Compared to other baselines, DarSwin achieves the best results (in terms of Top-1 accuracy) with significant gains when trained on bounded levels of distortions (very-low, low, medium, and high) and tested on all including out-of-distribution distortions. The code and models are publicly available at https://lvsn.github.io/darswin/
****
Fine-grained Unsupervised Domain Adaptation for Gait Recognition
Kang Ma, Ying Fu, Dezhi Zheng, Yunjie Peng, Chunshui Cao, Yongzhen Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11313-11322
Gait recognition has emerged as a promising technique for the long-range retrieval of pedestrians, providing numerous advantages such as accurate identification in challenging conditions and non-intrusiveness, making it highly desirable for improving public safety and security. However, the high cost of labeling datasets, which is a prerequisite for most existing fully supervised approaches, poses a significant obstacle to the development of gait recognition. Recently, some unsupervised methods for gait recognition have shown promising results. However,

these methods mainly rely on a fine-tuning approach that does not sufficiently consider the relationship between source and target domains, leading to the catastrophic forgetting of source domain knowledge. This paper presents a novel perspective that adjacent-view sequences exhibit overlapping views, which can be leveraged by the network to gradually attain cross-view and cross-dressing capabilities without pre-training on the labeled source domain. Specifically, we propose a fine-grained Unsupervised Domain Adaptation (UDA) framework that iteratively alternates between two stages. The initial stage involves offline clustering, which transfers knowledge from the labeled source domain to the unlabeled target domain and adaptively generates pseudo-labels according to the expressiveness of each part. Subsequently, the second stage encompasses online training, which further achieves cross-dressing capabilities by continuously learning to distinguish numerous features of source and target domains. The effectiveness of the proposed method is demonstrated through extensive experiments conducted on widely-used public gait datasets.
****

Cross-Modal Orthogonal High-Rank Augmentation for RGB-Event Transformer-Trackers
Zhiyu Zhu, Junhui Hou, Dapeng Oliver Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22045-22055
This paper addresses the problem of cross-modal object tracking from RGB videos and event data. Rather than constructing a complex cross-modal fusion network, we explore the great potential of a pre-trained vision Transformer (ViT). Particularly, we delicately investigate plug-and-play training augmentations that encourage the ViT to bridge the vast distribution gap between the two modalities, enabling comprehensive cross-modal information interaction and thus enhancing its ability. Specifically, we propose a mask modeling strategy that randomly masks a specific modality of some tokens to enforce the interaction between tokens from different modalities interacting proactively. To mitigate network oscillations resulting from the masking strategy and further amplify its positive effect, we then theoretically propose an orthogonal high-rank loss to regularize the attention matrix. Extensive experiments demonstrate that our plug-and-play training augmentation techniques can significantly boost state-of-the-art one-stream and two-stream trackers to a large extent in terms of both tracking precision and success rate. Our new perspective and findings will potentially bring insights to the field of leveraging powerful pre-trained ViTs to model cross-modal data. The code is publicly available at https://github.com/ZHU-Zhiyu/High-Rank_RGB-Event_Tracker.
****

Take-A-Photo: 3D-to-2D Generative Pre-training of Point Cloud Models
Ziyi Wang, Xumin Yu, Yongming Rao, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5640-5650
With the overwhelming trend of mask image modeling led by MAE, generative pre-training has shown a remarkable potential to boost the performance of fundamental models in 2D vision. However, in 3D vision, the over-reliance on Transformer-based backbones and the unordered nature of point clouds have restricted the further development of gen- erative pre-training. In this paper, we propose a novel 3D-to- 2D generative pre-training method that is adaptable to any point cloud model. We propose to generate view images from different instructed poses via the cross-attention mechanism as the pre-training scheme. Generating view images has more precise supervision than its point cloud counterpart, thus assisting 3D backbones to have a finer comprehension of the geometrical structure and stereoscopic relations of the point cloud. Experimental results have proved the su- periority of our proposed 3D-to-2D generative pre-training over previous pre-training methods. Our method is also ef- fective in boosting the performance of architecture-oriented approaches, achieving state-of-the-art performance when fine-tuning on ScanObjectNN classification and ShapeNet- Part segmentation tasks. Code is available at https: //github.com/wangzy22/TakeAPhoto.
****

Open-vocabulary Panoptic Segmentation with Embedding Modulation
Xi Chen, Shuang Li, Ser-Nam Lim, Antonio Torralba, Hengshuang Zhao; Proceedings

of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11
41-1150
Open-vocabulary segmentation is attracting increasing attention due to its criti
cal applications in the real world. Traditional closed-vocabulary segmentation m
ethods are not able to characterize novel objects, whereas several recent open-v
ocabulary attempts obtain unsatisfactory results, i.e., notable performance redu
ction on the closed-vocabulary and massive demand for extra training data. To th
is end, we propose OPSNet, an omnipotent and data-efficient framework for Open-v
ocabulary Panoptic Segmentation. Specifically, the exquisitely designed Embeddin
g Modulation module, together with several meticulous components, enables adequa
te embedding enhancement and information exchange between the segmentation backb
one and the visual-linguistic well-aligned CLIP encoder, resulting in superior s
egmentation performance under both open- and closed vocabulary settings and much
 fewer need of additional data. Extensive experimental evaluations are conducted
 across multiple datasets(e.g., COCO, ADE20K, Cityscapes, and PascalContext) und
er various circumstances, where the proposed OPSNet achieves state-of-the-art re
sults, which demonstrates the effectiveness and generality of the proposed appro
ach. The code and trained models will be made publicly available.
****
Beyond Single Path Integrated Gradients for Reliable Input Attribution via Rando
mized Path Sampling
Giyoung Jeon, Haedong Jeong, Jaesik Choi; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2023, pp. 2052-2061
Input attribution is a widely used explanation method for deep neural networks,
especially in visual tasks. Among various attribution methods, Integrated Gradie
nts (IG) is frequently used because of its model-agnostic applicability and desi
rable axioms. However, previous work has shown that such method often produces n
oisy and unreliable attributions during the integration of the gradients over th
e path defined in the input space. In this paper, we tackle this issue by estima
ting the distribution of the possible attributions according to the integrating
path selection. We show that such noisy attribution can be reduced by aggregatin
g attributions from the multiple paths instead of using a single path. Inspired
by Stick-Breaking Process (SBP), we suggest a random process to generate rich an
d various sampling of the gradient integrating path. Using multiple input attrib
utions obtained from randomized path, we propose a novel attribution measure usi
ng the distribution of attributions at each input features. We identify proposed
 method qualitatively show less-noisy and object-aligned attribution and its fea
sibility through the quantitative evaluations.
****
Instance-aware Dynamic Prompt Tuning for Pre-trained Point Cloud Models
Yaohua Zha, Jinpeng Wang, Tao Dai, Bin Chen, Zhi Wang, Shu-Tao Xia; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14
161-14170
Pre-trained point cloud models have found extensive applications in 3D understan
ding tasks like object classification and part segmentation. However, the prevai
ling strategy of full fine-tuning in downstream tasks leads to large per-task st
orage overhead for model parameters, which limits the efficiency when applying l
arge-scale pre-trained models. Inspired by the recent success of visual prompt t
uning (VPT), this paper attempts to explore prompt tuning on pre-trained point c
loud models, to pursue an elegant balance between performance and parameter effi
ciency. We find while instance-agnostic static prompting, e.g. VPT, shows some e
fficacy in downstream transfer, it is vulnerable to the distribution diversity c
aused by various types of noises in real-world point cloud data. To conquer this
 limitation, we propose a novel Instance-aware Dynamic Prompt Tuning (IDPT) stra
tegy for pre-trained point cloud models. The essence of IDPT is to develop a dyn
amic prompt generation module to perceive semantic prior features of each point
cloud instance and generate adaptive prompt tokens to enhance the model's robust
ness. Notably, extensive experiments demonstrate that IDPT outperforms full fine
-tuning in most tasks with a mere 7% of the trainable parameters, providing a pr
omising solution to parameter-efficient learning for pre-trained point cloud mod

els. Code is available at https://github.com/zyh16143998882/ICCV23-IDPT.
****
How to Boost Face Recognition with StyleGAN?
Artem Sevastopolskiy, Yury Malkov, Nikita Durasov, Luisa Verdoliva, Matthias Nießner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20924-20934
State-of-the-art face recognition systems require huge amounts of labeled training data. Given the priority of privacy in face recognition applications, the data is limited to celebrity web crawls, which have issues such as skewed distributions of ethnicities and limited numbers of identities. On the other hand, the self-supervised revolution in the industry motivates research on adaptation of the related techniques to facial recognition. One of the most popular practical tricks is to augment the dataset by the samples drawn from the high-resolution high-fidelity models (e.g. StyleGAN-like), while preserving the identity. We show that a simple approach based on fine-tuning an encoder for StyleGAN allows to improve upon the state-of-the-art facial recognition and performs better compared to training on synthetic face identities. We also collect large-scale unlabeled datasets with controllable ethnic constitution -- AfricanFaceSet-5M (5 million images of different people) and AsianFaceSet-3M (3 million images of different people) and we show that pretraining on each of them improves recognition of the respective ethnicities (as well as also others), while combining all unlabeled datasets results in the biggest performance increase. Our self-supervised strategy is the most useful with limited amounts of labeled training data, which can be beneficial for more tailored face recognition tasks and when facing privacy concerns. Evaluation is provided based on a standard RFW dataset and a new large-scale RB-WebFace benchmark.
****
Text2Tex: Text-driven Texture Synthesis via Diffusion Models
Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, Matthias Nießner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18558-18568
We present Text2Tex, a novel method for generating high-quality textures for 3D meshes from the given text prompts. Our method incorporates inpainting into a pre-trained depth-aware image diffusion model to progressively synthesize high resolution partial textures from multiple viewpoints. To avoid accumulating inconsistent and stretched artifacts across views, we dynamically segment the rendered view into a generation mask, which represents the generation status of each visible texel. This partitioned view representation guides the depth-aware inpainting model to generate and update partial textures for the corresponding regions. Furthermore, we propose an automatic view sequence generation scheme to determine the next best view for updating the partial texture. Extensive experiments demonstrate that our method significantly outperforms the existing text-driven approaches and GAN-based methods.
****
MUVA: A New Large-Scale Benchmark for Multi-View Amodal Instance Segmentation in the Shopping Scenario
Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, Tiejun Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23504-23513
Amodal Instance Segmentation (AIS) endeavors to accurately deduce complete object shapes that are partially or fully occluded. However, the inherent ill-posed nature of single-view datasets poses challenges in determining occluded shapes. A multi-view framework may help alleviate this problem, as humans often adjust their perspective when encountering occluded objects. At present, this approach has not yet been explored by existing methods and datasets. To bridge this gap, we propose a new task called Multi-view Amodal Instance Segmentation (MAIS) and introduce the MUVA dataset, the first MUlti-View AIS dataset that takes the shopping scenario as instantiation. MUVA provides comprehensive annotations, including multi-view amodal/visible segmentation masks, 3D models, and depth maps, making it the largest image-level AIS dataset in terms of both the number of images an

d instances. Additionally, we propose a new method for aggregating representative features across different instances and views, which demonstrates promising results in accurately predicting occluded objects from one viewpoint by leveraging information from other viewpoints. Besides, we also demonstrate that MUVA can benefit the AIS task in real-world scenarios.
****
Foreground-Background Separation through Concept Distillation from Generative Image Foundation Models
Mischa Dombrowski, Hadrien Reynaud, Matthew Baugh, Bernhard Kainz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 988-998
Curating datasets for object segmentation is a difficult task. With the advent of large-scale pre-trained generative models, conditional image generation has been given a significant boost in result quality and ease of use. In this paper, we present a novel method that enables the generation of general foreground-background segmentation models from simple textual descriptions, without requiring segmentation labels. We leverage and explore pre-trained latent diffusion models, to automatically generate weak segmentation masks for concepts and objects. The masks are then used to fine-tune the diffusion model on an inpainting task, which enables fine-grained removal of the object, while at the same time providing a synthetic foreground and background dataset. We demonstrate that using this method beats previous methods in both discriminative and generative performance and closes the gap with fully supervised training while requiring no pixel-wise object labels. We show results on the task of segmenting four different objects (humans, dogs, cars, birds) and a use case scenario in medical image analysis. The code is available at https://github.com/MischaD/fobadiffusion.
****
ENVIDR: Implicit Differentiable Renderer with Neural Environment Lighting
Ruofan Liang, Huiting Chen, Chunlin Li, Fan Chen, Selvakumar Panneer, Nandita Vijaykumar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 79-89
Recent advances in neural rendering have shown great potential for reconstructing scenes from multiview images. However, accurately representing objects with glossy surfaces remains a challenge for existing methods. In this work, we introduce ENVIDR, a rendering and modeling framework for high-quality rendering and reconstruction of surfaces with challenging specular reflections. To achieve this, we first propose a novel neural renderer with decomposed rendering components to learn the interaction between surface and environment lighting. This renderer is trained using existing physically based renderers and is decoupled from actual scene representations. We then propose an SDF-based neural surface model that leverages this learned neural renderer to represent general scenes. Our model additionally synthesizes indirect illuminations caused by inter-reflections from shiny surfaces by marching surface-reflected rays. We demonstrate that our method outperforms state-of-art methods on challenging shiny scenes, providing high-quality rendering of specular reflections while also enabling material editing and scene relighting.
****
Not All Steps are Created Equal: Selective Diffusion Distillation for Image Manipulation
Luozhou Wang, Shuai Yang, Shu Liu, Ying-cong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7472-7481
Conditional diffusion models have demonstrated impressive performance in image manipulation tasks. The general pipeline involves adding noise to the image and then denoising it. However, this method faces a trade-off problem: adding too much noise affects the fidelity of the image while adding too little affects its editability. This largely limits their practical applicability. In this paper, we propose a novel framework, Selective Diffusion Distillation (SDD), that ensures both the fidelity and editability of images. Instead of directly editing images with a diffusion model, we train a feedforward image manipulation network under the guidance of the diffusion model. Besides, we propose an effective indicator

to select the semantic-related timestep to obtain the correct semantic guidance from the diffusion model. This approach successfully avoids the dilemma caused by the diffusion process. Our extensive experiments demonstrate the advantages of our framework.
****
SeiT: Storage-Efficient Vision Training with Tokens Using 1% of Pixel Storage
Song Park, Sanghyuk Chun, Byeongho Heo, Wonjae Kim, Sangdoo Yun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17248-17259
We need billion-scale images to achieve more generalizable and ground-breaking vision models, as well as massive dataset storage to ship the images (e.g., the LAION-4B dataset needs 240TB storage space). However, it has become challenging to deal with unlimited dataset storage with limited storage infrastructure. A number of storage-efficient training methods have been proposed to tackle the problem, but they are rarely scalable or suffer from severe damage to performance. In this paper, we propose a storage-efficient training strategy for vision classifiers for large-scale datasets (e.g., ImageNet) that only uses 1024 tokens per instance without using the raw level pixels; our token storage only needs <1% of the original JPEG-compressed raw pixels. We also propose token augmentations and a Stem-adaptor module to make our approach able to use the same architecture as pixel-based approaches with only minimal modifications on the stem layer and the carefully tuned optimization settings. Our experimental results on ImageNet-1K show that our method significantly outperforms other storage-efficient training methods with a large gap. We further show the effectiveness of our method in other practical scenarios, storage-efficient pre-training, and continual learning. We will make our implementation and tokenized dataset publicly after the acceptance.
****
ALIP: Adaptive Language-Image Pre-Training with Synthetic Caption
Kaicheng Yang, Jiankang Deng, Xiang An, Jiawei Li, Ziyong Feng, Jia Guo, Jing Yang, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2922-2931
Contrastive Language-Image Pre-training (CLIP) has significantly boosted the performance of various vision-language tasks by scaling up the dataset with image-text pairs collected from the web. However, the presence of intrinsic noise and unmatched image-text pairs in web data can potentially affect the performance of representation learning. To address this issue, we first utilize the OFA model to generate synthetic captions that focus on the image content. The generated captions contain complementary information that is beneficial for pre-training. Then, we propose an Adaptive Language-Image Pre-training (ALIP), a bi-path model that integrates supervision from both raw text and synthetic caption. As the core components of ALIP, the Language Consistency Gate (LCG) and Description Consistency Gate (DCG) dynamically adjust the weights of samples and image-text/caption pairs during the training process. Meanwhile, the adaptive contrastive loss can effectively reduce the impact of noise data and enhances the efficiency of pre-training data. We validate ALIP with experiments on different scales of models and pre-training datasets. Experiments results show that ALIP achieves state-of-the-art performance on multiple downstream tasks including zero-shot image-text retrieval and linear probe. To facilitate future research, the code and pre-trained models are released at https://github.com/deepglint/ALIP.
****
GeoUDF: Surface Reconstruction from 3D Point Clouds via Geometry-guided Distance Representation
Siyu Ren, Junhui Hou, Xiaodong Chen, Ying He, Wenping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14214-14224
We present a learning-based method, namely GeoUDF, to tackle the long-standing and challenging problem of reconstructing a discrete surface from a sparse point cloud. To be specific, we propose a geometry-guided learning method for UDF and its gradient estimation that explicitly formulates the unsigned distance of a qu

ery point as the learnable affine averaging of its distances to the tangent planes of neighboring points on the surface. Besides, we model the local geometric structure of the input point clouds by explicitly learning a quadratic polynomial for each point. This not only facilitates upsampling the input sparse point cloud but also naturally induces unoriented normal, which further augments UDF estimation. Finally, to extract triangle meshes from the predicted UDF, we propose a customized edge-based marching cube module. We conduct extensive experiments and ablation studies to demonstrate the significant advantages of our method over state-of-the-art methods in terms of reconstruction accuracy, efficiency, and generality. The source code is publicly available at https://github.com/rsy6318/GeoUDF.
****
LaPE: Layer-adaptive Position Embedding for Vision Transformers with Independent Layer Normalization
Runyi Yu, Zhennan Wang, Yinhuai Wang, Kehan Li, Chang Liu, Haoyi Duan, Xiangyang Ji, Jie Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5886-5896
Position information is critical for Vision Transformers (VTs) due to the permutation-invariance of self-attention operations. A typical way to introduce position information is adding the absolute Position Embedding (PE) to patch embedding before entering VTs. However, this approach operates the same Layer Normalization (LN) to token embedding and PE, and delivers the same PE to each layer. This results in restricted and monotonic PE across layers, as the shared LN affine parameters are not dedicated to PE, and the PE cannot be adjusted on a per-layer basis. To overcome these limitations, we propose using two independent LNs for token embeddings and PE in each layer, and progressively delivering PE across layers. By implementing this approach, VTs will receive layer-adaptive and hierarchical PE. We name our method as Layer-adaptive Position Embedding, abbreviated as LaPE, which is simple, effective, and robust. Extensive experiments on image classification, object detection, and semantic segmentation demonstrate that LaPE significantly outperforms the default PE method. For example, LaPE improves +1.06% for CCT on CIFAR100, +1.57% for DeiT-Ti on ImageNet-1K, +0.7 box AP and +0.5 mask AP for ViT-Adapter-Ti on COCO, and +1.37 mIoU for tiny Segmenter on ADE20K. This is remarkable considering LaPE only increases negligible parameters, memory, and computational cost.
****
CLIP2Point: Transfer CLIP to Point Cloud Classification with Image-Depth Pre-Training
Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson W.H. Lau, Wanli Ouyang, Wangmeng Zuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22157-22167
Pre-training across 3D vision and language remains under development because of limited training data. Recent works attempt to transfer vision-language (V-L) pre-training methods to 3D vision. However, the domain gap between 3D and images is unsolved, so that V-L pre-trained models are restricted in 3D downstream tasks. To address this issue, we propose CLIP2Point, an image-depth pre-training method by contrastive learning to transfer CLIP to the 3D domain, and adapt it to point cloud classification. We introduce a new depth rendering setting that forms a better visual effect, and then render 52,460 pairs of images and depth maps from ShapeNet for pre-training. The pre-training scheme of CLIP2Point combines cross-modality learning to enforce the depth features for capturing expressive visual and textual features and intra-modality learning to enhance the invariance of depth aggregation. Additionally, we propose a novel Gated Dual-Path Adapter (GDPA), i.e., a dual-path structure with global-view aggregators and gated fusion for downstream representative learning. It allows the ensemble of CLIP and CLIP2Point, tuning pre-training knowledge to downstream tasks in an efficient adaptation. Experimental results show that CLIP2Point is effective in transferring CLIP knowledge to 3D vision. Our CLIP2Point outperforms other 3D transfer learning and pre-training networks, achieving state-of-the-art results on zero-shot, few-shot, and fully-supervised classification.

****
Parametric Classification for Generalized Category Discovery: A Baseline Study
Xin Wen, Bingchen Zhao, Xiaojuan Qi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16590-16600
Generalized Category Discovery (GCD) aims to discover novel categories in unlabelled datasets using knowledge learned from labelled samples. Previous studies argued that parametric classifiers are prone to overfitting to seen categories, and endorsed using a non-parametric classifier formed with semi-supervised k-means. However, in this study, we investigate the failure of parametric classifiers, verify the effectiveness of previous design choices when high-quality supervision is available, and identify unreliable pseudo-labels as a key problem. We demonstrate that two prediction biases exist: the classifier tends to predict seen classes more often, and produces an imbalanced distribution across seen and novel categories. Based on these findings, we propose a simple yet effective parametric classification method that benefits from entropy regularisation, achieves state-of-the-art performance on multiple GCD benchmarks and shows strong robustness to unknown class numbers. We hope the investigation and proposed simple framework can serve as a strong baseline to facilitate future studies in this field. Our code is available at: https://github.com/CVMI-Lab/SimGCD.
****
MeMOTR: Long-Term Memory-Augmented Transformer for Multi-Object Tracking
Ruopeng Gao, Limin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9901-9910
As a video task, Multiple Object Tracking (MOT) is expected to capture temporal information of targets effectively. Unfortunately, most existing methods only explicitly exploit the object features between adjacent frames, while lacking the capacity to model long-term temporal information. In this paper, we propose MeMOTR, a long-term memory-augmented Transformer for multi-object tracking. Our method is able to make the same object's track embedding more stable and distinguishable by leveraging long-term memory injection with a customized memory-attention layer. This significantly improves the target association ability of our model. Experimental results on DanceTrack show that MeMOTR impressively surpasses the state-of-the-art method by 7.9% and 13.0% on HOTA and AssA metrics, respectively. Furthermore, our model also outperforms other Transformer-based methods on association performance on MOT17 and generalizes well on BDD100K. Code is available at https://github.com/MCG-NJU/MeMOTR.
****
RawHDR: High Dynamic Range Image Reconstruction from a Single Raw Image
Yunhao Zou, Chenggang Yan, Ying Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12334-12344
High dynamic range (HDR) images can record much more intensity levels than usual ones. Existing methods mainly reconstruct HDR images from the 8-bit low dynamic range (LDR) sRGB images that have been degraded by the camera processing pipeline. However, recovering extremely high dynamic range scenes from such low bit-depth data is challenging. Unlike existing methods, the core idea of this work is to incorporate more informative Raw sensor data to generate HDR images, aiming to recover scene information in hard regions (the darkest and brightest areas of an HDR scene). We propose a model customized for Raw images, considering the unique feature of Raw data to learn the Raw-to-HDR mapping. Specifically, we learn exposure masks to separate the hard and easy regions of a high dynamic scene. Then, we introduce two important guidances, dual intensity guidance, which guides less informative channels with more informative ones, and global spatial guidance which hallucinates scene details from a longer spatial range. To verify our Raw-to-HDR approach, we collect a large and high-quality Raw/HDR paired dataset for both training and testing, which will be made available publicly. We verify the superiority of the proposed Raw-to-HDR reconstruction model, as well as our newly captured dataset in the experiments.
****
Denoising Diffusion Autoencoders are Unified Self-supervised Learners
Weilai Xiang, Hongyu Yang, Di Huang, Yunhong Wang; Proceedings of the IEEE/CVF I

nternational Conference on Computer Vision (ICCV), 2023, pp. 15802-15812
Inspired by recent advances in diffusion models, which are reminiscent of denois ing autoencoders, we investigate whether they can acquire discriminative represe ntations for classification via generative pre-training. This paper shows that t he networks in diffusion models, namely denoising diffusion autoencoders (DDAE), are unified self-supervised learners: by pre-training on unconditional image ge neration, DDAE has already learned strongly linear-separable representations wit hin its intermediate layers without auxiliary encoders, thus making diffusion pr e-training emerge as a general approach for generative-and-discriminative dual l earning. To validate this, we conduct linear probe and fine-tuning evaluations. Our diffusion-based approach achieves 95.9% and 50.0% linear evaluation accuraci es on CIFAR-10 and Tiny-ImageNet, respectively, and is comparable to contrastive learning and masked autoencoders for the first time. Transfer learning from Ima geNet also confirms the suitability of DDAE for Vision Transformers, suggesting the potential to scale DDAEs as unified foundation models. Code is available at github.com/FutureXiang/ddae.
****
Robust Object Modeling for Visual Tracking
Yidong Cai, Jie Liu, Jie Tang, Gangshan Wu; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2023, pp. 9589-9600
Object modeling has become a core part of recent tracking frameworks. Current po pular tackers use Transformer attention to extract the template feature separate ly or interactively with the search region. However, separate template learning lacks communication between the template and search regions, which brings diffic ulty in extracting discriminative target-oriented features. On the other hand, i nteractive template learning produces hybrid template features, which may introd uce potential distractors to the template via the cluttered search regions. To e njoy the merits of both methods, we propose a robust object modeling framework f or visual tracking (ROMTrack), which simultaneously models the inherent template and the hybrid template features. As a result, harmful distractors can be suppr essed by combining the inherent features of target objects with search regions' guidance. Target-related features can also be extracted using the hybrid templat e, thus resulting in a more robust object modeling framework. To further enhance robustness, we present novel variation tokens to depict the ever-changing appea rance of target objects. Variation tokens are adaptable to object deformation an d appearance variations, which can boost overall performance with negligible com putation. Experiments show that our ROMTrack sets a new state-of-the-art on mult iple benchmarks.
****
FSI: Frequency and Spatial Interactive Learning for Image Restoration in Under-D isplay Cameras
Chengxu Liu, Xuan Wang, Shuai Li, Yuzhi Wang, Xueming Qian; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12537-1254 6
Under-display camera (UDC) systems remove the screen notch for bezel-free displa ys and provide a better interactive experience. The main challenge is that the p ixel array of light-emitting diodes used for display diffracts and attenuates th e incident light, leading to complex degradation. Existing models eliminate spat ial diffraction by maximizing model capacity through complex design and ignore t he periodic distribution of diffraction in the frequency domain, which prevents these approaches from satisfactory results. In this paper, we introduce a new pe rspective to handle various diffraction in UDC images by jointly exploring the f eature restoration in the frequency and spatial domains, and present a Frequency and Spatial Interactive Learning Network (FSI). It consists of a series of well -designed Frequency-Spatial Joint (FSJ) modules for feature learning and a color transform module for color enhancement. In particular, in the FSJ module, a fre quency learning block uses the Fourier transform to eliminate spectral bias, a s patial learning block uses a multi-distillation structure to supplement the abse nce of local details, and a dual transfer unit to facilitate the interactive lea rning between features of different domains. Experimental results demonstrate th

e superiority of the proposed FSI over state-of-the-art models, through extensive quantitative and qualitative evaluations in three widely-used UDC benchmarks.

****

Cross-view Topology Based Consistent and Complementary Information for Deep Multi-view Clustering

Zhibin Dong, Siwei Wang, Jiaqi Jin, Xinwang Liu, En Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19440-19451

Multi-view clustering aims to extract valuable information from different sources or perspectives. Over the years, the deep neural network has demonstrated its superior representation learning capability in multi-view clustering and achieved impressive performance. However, most existing deep clustering approaches are dedicated to merging and exploring the consistent latent representation across multiple views while overlooking the abundant complementary information in each view. Furthermore, finding correlations between multiple views in an unsupervised setting is a significant challenge. To tackle these issues, we present a novel Cross-view Topology based Consistent and Complementary information extraction framework, termed CTCC. In detail, deep embedding can be obtained from the bipartite graph learning module for each view individually. CTCC then constructs the cross-view topological graph based on the OT distance between the bipartite graph of each view. Utilizing the above graph, we maximize the mutual information across views to learn consistent information and enhance the complementarity of each view by selectively isolating distributions from each other. Extensive experiments on five challenging datasets verify that CTCC outperforms existing methods significantly.

****

Distribution-Consistent Modal Recovering for Incomplete Multimodal Learning

Yuanzhi Wang, Zhen Cui, Yong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22025-22034

Recovering missed modality is popular in incomplete multimodal learning because it usually benefits downstream tasks. However, the existing methods often directly estimate missed modalities from the observed ones by deep neural networks, lacking consideration of the distribution gap between modalities, resulting in the inconsistency of distributions between the recovered data and true data. To mitigate this issue, in this work, we propose a novel recovery paradigm, Distribution-Consistent Modal Recovering (DiCMoR), to transfer the distributions from available modalities to missed modalities, which thus maintains the distribution consistency of recovered data. In particular, we design a class-specific flow based modality recovery method to transform cross-modal distributions on the condition of sample class, which could well predict a distribution-consistent space for missing modality by virtue of the invertibility and exact density estimation of normalizing flow. The generated data from the predicted distribution is jointly integrated with available modalities for the task of classification. Experiments demonstrate that DiCMoR gains superior performances and is more robust than existing state-of-the-art methods under various missing patterns. Visualization results show that the distribution gaps between recovered modalities and missing modalities are mitigated.

****

ContactGen: Generative Contact Modeling for Grasp Generation

Shaowei Liu, Yang Zhou, Jimei Yang, Saurabh Gupta, Shenlong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20609-20620

This paper presents a novel object-centric contact representation ContactGen for hand-object interaction. The ContactGen comprises 3 components: a contact map indicates the contact location, a part map represents the contact hand part, and a direction map tells the contact direction within each part. Given an input object, we propose a conditional generative model to predict ContactGen and adopt model-based optimization to predict diverse and geometrically feasible grasps. Experimental results demonstrate our method can generate high-fidelity and diverse human grasps for various objects.

****

Temporal Collection and Distribution for Referring Video Object Segmentation
Jiajin Tang, Ge Zheng, Sibei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15466-15476
Referring video object segmentation aims to segment a referent throughout a video sequence according to a natural language expression. It requires aligning the natural language expression with the objects' motions and their dynamic associations at the global video level but segmenting objects at the frame level. To achieve this goal, we propose to simultaneously maintain a global referent token and a sequence of object queries, where the former is responsible for capturing video-level referent according to the language expression, while the latter serves to better locate and segment objects with each frame. Furthermore, to explicitly capture object motions and spatial-temporal cross-modal reasoning over objects, we propose a novel temporal collection-distribution mechanism for interacting between the global referent token and object queries. Specifically, the temporal collection mechanism collects global information for the referent token from object queries to the temporal motions to the language expression. In turn, the temporal distribution first distributes the referent token to the referent sequence across all frames and then performs efficient cross-frame reasoning between the referent sequence and object queries in every frame. Experimental results show that our method outperforms state-of-the-art methods on all benchmarks consistently and significantly.
****
SA-BEV: Generating Semantic-Aware Bird's-Eye-View Feature for Multi-view 3D Object Detection
Jinqing Zhang, Yanan Zhang, Qingjie Liu, Yunhong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3348-3357
Recently, the pure camera-based Bird's-Eye-View (BEV) perception provides a feasible solution for economical autonomous driving. However, the existing BEV-based multi-view 3D detectors generally transform all image features into BEV features, without considering the problem that the large proportion of background information may submerge the object information. In this paper, we propose Semantic-Aware BEV Pooling (SA-BEVPool), which can filter out background information according to the semantic segmentation of image features and transform image features into semantic-aware BEV features. Accordingly, we propose BEV-Paste, an effective data augmentation strategy that closely matches with semantic-aware BEV feature. In addition, we design a Multi-Scale Cross-Task (MSCT) head, which combines task-specific and cross-task information to predict depth distribution and semantic segmentation more accurately, further improving the quality of semantic-aware BEV feature. Finally, we integrate the above modules into a novel multi-view 3D object detection framework, namely SA-BEV. Experiments on nuScenes show that SA-BEV achieves state-of-the-art performance. Code has been available at https://github.com/mengtan00/SA-BEV.git.
****
Variational Degeneration to Structural Refinement: A Unified Framework for Superimposed Image Decomposition
Wenyu Li, Yan Xu, Yang Yang, Haoran Ji, Yue Lang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12206-12216
Decomposing a single mixed image into individual image layers is the common crux of a classical category of tasks in image restoration. Several unified frameworks have been proposed that can handle different types of degradation in superimposed image decomposition. However, there are always undesired structural distortions in the separated images when dealing with complicated degradation patterns. In this paper, we propose a unified framework for superimposed image decomposition that can cope with intricate degradation patterns adaptively. Considering the different mixing patterns between the layers, we introduce a degeneration representation in the latent space to mine the intrinsic relationship between the superimposed image and the degeneration pattern. Moreover, by extracting structure-guided knowledge from the superimposed image, we further propose structural guidance refinement to avoid confusing content caused by structure distortion. Extensive experiments have demonstrated that our method remarkably outperforms other

popular image separation frameworks. The method also achieves competitive results on related applications including image deraining, image reflection removal, and image shadow removal, which validates the generalization of the framework.
****
Global Knowledge Calibration for Fast Open-Vocabulary Segmentation
Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, Yunchao Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 797-807
Recent advancements in pre-trained vision-language models, such as CLIP, have enabled the segmentation of arbitrary concepts solely from textual inputs, a process commonly referred to as open-vocabulary semantic segmentation (OVS). However, existing OVS techniques confront a fundamental challenge: the trained classifier tends to overfit on the base classes observed during training, resulting in suboptimal generalization performance to unseen classes. To mitigate this issue, recent studies have proposed the use of an additional frozen pre-trained CLIP for classification. Nonetheless, this approach incurs heavy computational overheads as the CLIP vision encoder must be repeatedly forward-passed for each mask, rendering it impractical for real-world applications. To address this challenge, our objective is to develop a fast OVS model that can perform comparably or better without the extra computational burden of the CLIP image encoder during inference. To this end, we propose a core idea of preserving the generalizable representation when fine-tuning on known classes. Specifically, we introduce a text diversification strategy that generates a set of synonyms for each training category, which prevents the learned representation from collapsing onto specific known category names. Additionally, we employ a text-guided knowledge distillation method to preserve the generalizable knowledge of CLIP. Extensive experiments demonstrate that our proposed model achieves robust generalization performance across various datasets. Furthermore, we perform a preliminary exploration of open-vocabulary video segmentation and present a benchmark that can facilitate future open-vocabulary research in the video domain.
****
Ego-Humans: An Ego-Centric 3D Multi-Human Benchmark
Rawal Khirodkar, Aayush Bansal, Lingni Ma, Richard Newcombe, Minh Vo, Kris Kitani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19807-19819
We present EgoHumans, a new multi-view multi-human video benchmark to advance the state-of-the-art of egocentric human 3D pose estimation and tracking. Existing egocentric benchmarks either capture single subject or indoor-only scenarios, which limit the generalization of computer vision algorithms for real-world applications. We propose a novel 3D capture setup to construct a comprehensive egocentric multi-human benchmark in the wild with annotations to support diverse tasks such as human detection, tracking, 2D/3D pose estimation, and mesh recovery. We leverage consumer-grade wearable camera-equipped glasses for the egocentric view, which enables us to capture dynamic activities like playing tennis, fencing, volleyball, etc. Furthermore, our multi-view setup generates accurate 3D ground truth even under severe or complete occlusion. The dataset consists of more than 125k egocentric images, spanning diverse scenes with a particular focus on challenging and unchoreographed multi-human activities and fast-moving egocentric views. We rigorously evaluate existing state-of-the-art methods and highlight their limitations in the egocentric scenario, specifically on multi-human tracking. To address such limitations, we propose EgoFormer, a novel approach with a multi-stream transformer architecture and explicit 3D spatial reasoning to estimate and track the human pose. EgoFormer significantly outperforms prior art by 13.6% IDF1 on the EgoHumans dataset.
****
Focal Network for Image Restoration
Yuning Cui, Wenqi Ren, Xiaochun Cao, Alois Knoll; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13001-13011
Image restoration aims to reconstruct a sharp image from its degraded counterpart, which plays an important role in many fields. Recently, Transformer models ha

ve achieved promising performance on various image restoration tasks. However, their quadratic complexity remains an intractable issue for practical applications. The aim of this study is to develop an efficient and effective framework for image restoration. Inspired by the fact that different regions in a corrupted image always undergo degradations in various degrees, we propose to focus more on the important areas for reconstruction. To this end, we introduce a dual-domain selection mechanism to emphasize crucial information for restoration, such as edge signals and hard regions. In addition, we split high-resolution features to insert multi-scale receptive fields into the network, which improves both efficiency and performance. Finally, the proposed network, dubbed FocalNet, is built by incorporating these designs into a U-shaped backbone. Extensive experiments demonstrate that our model achieves state-of-the-art performance on ten datasets for three tasks, including single-image defocus deblurring, image dehazing, and image desnowing. Our code is available at https://github.com/c-yn/FocalNet.
****
Indoor Depth Recovery Based on Deep Unfolding with Non-Local Prior
Yuhui Dai, Junkang Zhang, Faming Fang, Guixu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12355-12364
In recent years, depth recovery based on deep networks has achieved great success. However, the existing state-of-the-art network designs perform like black boxes in depth recovery tasks, lacking a clear mechanism. Utilizing the property that there is a large amount of non-local common characteristics in depth images, we propose a novel model-guided depth recovery method, namely the DC-NLAR model. A non-local auto-regressive regular term is also embedded into our model to capture more non-local depth information. To fully use the excellent performance of neural networks, we develop a deep image prior to better describe the characteristic of depth images. We also introduce an implicit data consistency term to tackle the degenerate operator with high heterogeneity. We then unfold the proposed model into networks by using the half-quadratic splitting algorithm. This proposed method is experimented on the NYU-Depth V2 and SUN RGB-D datasets, and the experimental results achieve comparable performance to that of deep learning methods.
****
Compatibility of Fundamental Matrices for Complete Viewing Graphs
Martin Bråtelund, Felix Rydell; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3328-3336
This paper studies the problem of recovering cameras from a set of fundamental matrices. A set of fundamental matrices is said to be compatible if a set of cameras exists for which they are the fundamental matrices. We focus on the complete graph, where fundamental matrices for each pair of cameras are given. Previous work has established necessary and sufficient conditions for compatibility as rank and eigenvalue conditions on the n-view fundamental matrix obtained by concatenating the individual fundamental matrices. In this work, we show that the eigenvalue condition is redundant in the generic and collinear cases. We provide explicit homogeneous polynomials that describe necessary and sufficient conditions for compatibility in terms of the fundamental matrices and their epipoles. In this direction, we find that quadruple-wise compatibility is enough to ensure global compatibility for any number of cameras. We demonstrate that for four cameras, compatibility is generically described by triple-wise conditions and one additional equation involving all fundamental matrices.
****
GAFlow: Incorporating Gaussian Attention into Optical Flow
Ao Luo, Fan Yang, Xin Li, Lang Nie, Chunyu Lin, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9642-9651
Optical flow, or the estimation of motion fields from image sequences, is one of the fundamental problems in computer vision. Unlike most pixel-wise tasks that aim at achieving consistent representations of the same category, optical flow raises extra demands for obtaining local discrimination and smoothness, which yet is not fully explored by existing approaches. In this paper, we push Gaussian A

ttention (GA) into the optical flow models to accentuate local properties during representation learning and enforce the motion affinity during matching. Specifically, we introduce a novel Gaussian-Constrained Layer (GCL) which can be easily plugged into existing Transformer blocks to highlight the local neighborhood that contains fine-grained structural information. Moreover, for reliable motion analysis, we provide a new Gaussian-Guided Attention Module (GGAM) which not only inherits properties from Gaussian distribution to instinctively revolve around the neighbor fields of each point but also is empowered to put the emphasis on contextually related regions during matching. Our fully-equipped model, namely Gaussian Attention Flow network (GAFlow), naturally incorporates a series of novel Gaussian-based modules into the conventional optical flow framework for reliable motion analysis. Extensive experiments on standard optical flow datasets consistently demonstrate the exceptional performance of the proposed approach in terms of both generalization ability evaluation and online benchmark testing. Code is available at https://github.com/LA30/GAFlow.
****
MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge
Wei Lin, Leonid Karlinsky, Nina Shvetsova, Horst Possegger, Mateusz Kozinski, Rameswar Panda, Rogerio Feris, Hilde Kuehne, Horst Bischof; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2851-2862
Large scale Vision-Language (VL) models have shown tremendous success in aligning representations between visual and text modalities. This enables remarkable progress in zero-shot recognition, image generation & editing, and many other exciting tasks. However, VL models tend to over-represent objects while paying much less attention to verbs, and require additional tuning on video data for best zero-shot action recognition performance. While previous work relied on large-scale, fully-annotated data, in this work we propose an unsupervised approach. We adapt a VL model for zero-shot and few-shot action recognition using a collection of unlabeled videos and an unpaired action dictionary. Based on that, we leverage Large Language Models and VL models to build a text bag for each unlabeled video via matching, text expansion and captioning. We use those bags in a Multiple Instance Learning setup to adapt an image-text backbone to video data. Although finetuned on unlabeled video data, our resulting models demonstrate high transferability to numerous unseen zero-shot downstream tasks, improving the base VL model performance by up to 14%, and even comparing favorably to fully-supervised baselines in both zero-shot and few-shot video recognition transfer. The code is provided in supplementary and will be released upon acceptance.
****
Space Engage: Collaborative Space Supervision for Contrastive-Based Semi-Supervised Semantic Segmentation
Changqi Wang, Haoyu Xie, Yuhui Yuan, Chong Fu, Xiangyu Yue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 931-942
Semi-Supervised Semantic Segmentation (S4) aims to train a segmentation model with limited labeled images and a substantial volume of unlabeled images. To improve the robustness of representations, powerful methods introduce a pixel-wise contrastive learning approach in latent space (i.e., representation space) that aggregates the representations to their prototypes in a fully supervised manner. However, previous contrastive-based S4 methods merely rely on the supervision from the model's output (logits) in logit space during unlabeled training. In contrast, we utilize the outputs in both logit space and representation space to obtain supervision in a collaborative way. The supervision from two spaces plays two roles: 1) reduces the risk of over-fitting to incorrect semantic information in logits with the help of representations; 2) enhances the knowledge exchange between the two spaces. Furthermore, unlike previous approaches, we use the similarity between representations and prototypes as a new indicator to tilt training those under-performing representations and achieve a more efficient contrastive learning process. Results on two public benchmarks demonstrate the competitive performance of our method compared with state-of-the-art methods.
****

Delving into Motion-Aware Matching for Monocular 3D Object Tracking
Kuan-Chih Huang, Ming-Hsuan Yang, Yi-Hsuan Tsai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6909-6918
Recent advances of monocular 3D object detection facilitate the 3D multi-object tracking task based on low-cost camera sensors. In this paper, we find that the motion cue of objects along different time frames is critical in 3D multi-object tracking, which is less explored in existing monocular-based approaches. To this end, we propose MoMA-M3T, a framework that mainly consists of three motion-aware components. First, we represent the possible movement of an object related to all object tracklets in the feature space as its motion features. Then, we further model the historical object tracklet along the time frame in a spatial-temporal perspective via a motion transformer. Finally, we propose a motion-aware matching module to associate historical object tracklets and current observations as final tracking results. We conduct extensive experiments on the nuScenes and KITTI datasets to demonstrate that our MoMA-M3T achieves competitive performance against state-of-the-art methods. Moreover, the proposed tracker is flexible and can be easily plugged into existing image-based 3D object detectors without re-training. Code and models are available at https://github.com/kuanchihhuang/MoMA-M3T.
****
SoDaCam: Software-defined Cameras via Single-Photon Imaging
Varun Sundar, Andrei Ardelean, Tristan Swedish, Claudio Bruschini, Edoardo Charbon, Mohit Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8165-8176
Reinterpretable cameras are defined by their post-processing capabilities that exceed traditional imaging. We present "SoDaCam" that provides reinterpretable cameras at the granularity of photons, from photon-cubes acquired by single-photon devices. Photon-cubes represent the spatio-temporal detections of photons as a sequence of binary frames, at frame-rates as high as 100 kHz. We show that simple transformations of the photon-cube, or photon-cube projections, provide the functionality of numerous imaging systems including: exposure bracketing, flutter shutter cameras, video compressive systems, event cameras, and even cameras that move during exposure. Our photon-cube projections offer the flexibility of being software-defined constructs that are only limited by what is computable, and shot-noise. We exploit this flexibility to provide new capabilities for the emulated cameras. As an added benefit, our projections provide camera-dependent compression of photon-cubes, which we demonstrate using an implementation of our projections on a novel compute architecture that is designed for single-photon imaging.
****
Reference-guided Controllable Inpainting of Neural Radiance Fields
Ashkan Mirzaei, Tristan Aumentado-Armstrong, Marcus A. Brubaker, Jonathan Kelly, Alex Levinshtein, Konstantinos G. Derpanis, Igor Gilitschenski; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17815-17825
The popularity of Neural Radiance Fields (NeRFs) for view synthesis has led to a desire for NeRF editing tools. Here, we focus on inpainting regions in a view-consistent and controllable manner. In addition to the typical NeRF inputs and masks delineating the unwanted region in each view, we require only a single inpainted view of the scene, i.e., a reference view. We use monocular depth estimators to back-project the inpainted view to the correct 3D positions. Then, via a novel rendering technique, a bilateral solver can construct view-dependent effects in non-reference views, making the inpainted region appear consistent from any view. For non-reference disoccluded regions, which cannot be supervised by the single reference view, we devise a method based on image inpainters to guide both the geometry and appearance. Our approach shows superior performance to NeRF inpainting baselines, with the additional advantage that a user can control the generated scene via a single inpainted image.
****
Diffusion-Guided Reconstruction of Everyday Hand-Object Interaction Clips

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, Shubham Tulsiani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19717-19728
We tackle the task of reconstructing hand-object interactions from short video clips. Given an input video, our approach casts 3D inference as a per-video optimization and recovers a neural 3D representation of the object shape, as well as the time-varying motion and hand articulation. While the input video naturally provides some multi-view cues to guide 3D inference, these are insufficient on their own due to occlusions and limited viewpoint variations. To obtain accurate 3D, we augment the multi-view signals with generic data-driven priors to guide reconstruction. Specifically, we learn a diffusion network to model the conditional distribution of (geometric) renderings of objects conditioned on hand configuration and category label, and leverage it as a prior to guide the novel-view renderings of the reconstructed scene. We empirically evaluate our approach on egocentric videos across 6 object categories, and observe significant improvements over prior single-view and multi-view methods. Finally, we demonstrate our system's ability to reconstruct arbitrary clips from YouTube, showing both 1st and 3rd person interactions.
****
Decoupled Iterative Refinement Framework for Interacting Hands Reconstruction from a Single RGB Image
Pengfei Ren, Chao Wen, Xiaozheng Zheng, Zhou Xue, Haifeng Sun, Qi Qi, Jingyu Wang, Jianxin Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8014-8025
Reconstructing interacting hands from a single RGB image is a very challenging task. On the one hand, severe mutual occlusion and similar local appearance between two hands confuse the extraction of visual features, resulting in the misalignment of estimated hand meshes and the image. On the other hand, there are complex spatial relationship between interacting hands, which significantly increases the solution space of hand poses and increases the difficulty of network learning. In this paper, we propose a decoupled iterative refinement framework to achieve pixel-alignment hand reconstruction while efficiently modeling the spatial relationship between hands. Specifically, we define two feature spaces with different characteristics, namely 2D visual feature space and 3D joint feature space. First, we obtain joint-wise features from the visual feature map and utilize a graph convolution network and a transformer to perform intra- and inter-hand information interaction in the 3D joint feature space, respectively. Then, we project the joint features with global information back into the 2D visual feature space in an obfuscation-free manner and utilize the 2D convolution for pixel-wise enhancement. By performing multiple alternate enhancements in the two feature spaces, our method can achieve an accurate and robust reconstruction of interacting hands. Our method outperforms all existing two-hand reconstruction methods by a large margin on the InterHand2.6M dataset.
****
Fast Adversarial Training with Smooth Convergence
Mengnan Zhao, Lihe Zhang, Yuqiu Kong, Baocai Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4720-4729
Fast adversarial training (FAT) is beneficial for improving the adversarial robustness of neural networks. However, previous FAT work has encountered a significant issue known as catastrophic overfitting when dealing with large perturbation budgets, i.e. the adversarial robustness of models declines to near zero during training. To address this, we analyze the training process of prior FAT work and observe that catastrophic overfitting is accompanied by the appearance of loss convergence outliers. Therefore, we argue a moderately smooth loss convergence process will be a stable FAT process that solves catastrophic overfitting. To obtain a smooth loss convergence process, we propose a novel oscillatory constraint (dubbed ConvergeSmooth) to limit the loss difference between adjacent epochs. The convergence stride of ConvergeSmooth is introduced to balance convergence and smoothing. Likewise, we design weight centralization without introducing additional hyperparameters other than the loss balance coefficient. Our proposed methods are attack-agnostic and thus can improve the training stability of various F

AT techniques. Extensive experiments on popular datasets show that the proposed methods efficiently avoid catastrophic overfitting and outperform all previous F AT methods. Code is available at https://github.com/FAT-CS/ConvergeSmooth.
****

Who Are You Referring To? Coreference Resolution In Image Narrations
Arushi Goel, Basura Fernando, Frank Keller, Hakan Bilen; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2023, pp. 15247-15258
Coreference resolution aims to identify words and phrases which refer to the same entity in a text, a core task in natural language processing. In this paper, we extend this task to resolving coreferences in long-form narrations of visual scenes. First, we introduce a new dataset with annotated coreference chains and their bounding boxes, as most existing image-text datasets only contain short sentences without coreferring expressions or labeled chains. We propose a new technique that learns to identify coreference chains using weak supervision, only from image-text pairs and a regularization using prior linguistic knowledge. Our model yields large performance gains over several strong baselines in resolving coreferences.
We also show that coreference resolution helps improve grounding narratives in images.
****

DVGaze: Dual-View Gaze Estimation
Yihua Cheng, Feng Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20632-20641
Gaze estimation methods estimate gaze from facial appearance with a single camera. However, due to the limited view of a single camera, the captured facial appearance cannot provide complete facial information and thus complicate the gaze estimation problem. Recently, camera devices are rapidly updated. Dual cameras are affordable for users and have been applied in many devices.This development suggests us to further improve gaze estimation performance with dual-view gaze estimation. In this paper, we propose a dual-view gaze estimation network (DV-Gaze). DV-Gaze estimates dual-view gaze directions from a pair of images. We first propose a dual-view interactive convolution (DIC) block in DV-Gaze. DIC blocks exchange dual-view information during convolution in multiple feature scales. It fuses dual-view features along epipolar lines and compensate original features with the fused feature. We further propose a dual-view transformer to estimate gaze from dual-view features. Camera poses are encoded to indicate the position information in the transformer. We also consider the geometric relation between dual-view gaze directions and propose a dual-view gaze consistency loss for DV-Gaze. DV-Gaze achieves state-of-the-art performance on ETH-XGaze and EVE datasets. Our experiments also prove the potential of dual-view gaze estimation. We release codes in https://github.com/yihuacheng/DVGaze.
****

Dynamic Hyperbolic Attention Network for Fine Hand-object Reconstruction
Zhiying Leng, Shun-Cheng Wu, Mahdi Saleh, Antonio Montanaro, Hao Yu, Yin Wang, Nassir Navab, Xiaohui Liang, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14894-14904
Reconstructing both objects and hands in 3D from a single RGB image is complex. Existing methods rely on manually defined hand-object constraints in Euclidean space, leading to suboptimal feature learning. Compared with Euclidean space, hyperbolic space better preserves the geometric properties of meshes thanks to its exponentially-growing space distance, which amplifies the differences between the features based on similarity. In this work, we propose the first precise hand-object reconstruction method in hyperbolic space, namely Dynamic Hyperbolic Attention Network (DHANet), which leverages intrinsic properties of hyperbolic space to learn representative features. Our method that projects mesh and image features into a unified hyperbolic space includes two modules, i.e. dynamic hyperbolic graph convolution and image-attention hyperbolic graph convolution. With these two modules, our method learns mesh features with rich geometry-image multi-modal information and models better hand-object interaction. Our method provides a promising alternative for fine hand-object reconstruction in hyperbolic space. E

xtensive experiments on three public datasets demonstrate that our method outper
forms most state-of-the-art methods.
****

A-STAR: Test-time Attention Segregation and Retention for Text-to-image Synthesi
s
Aishwarya Agarwal, Srikrishna Karanam, K J Joseph, Apoorv Saxena, Koustava Goswa
mi, Balaji Vasan Srinivasan; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2023, pp. 2283-2293
While recent developments in text-to-image generative models have led to a suite
 of high-performing methods capable of producing creative imagery from free-form
 text, there are several limitations. By analyzing the cross-attention represent
ations of these models, we notice two key issues. First, for text prompts that c
ontain multiple concepts, there is a significant amount of pixel-space overlap (
i.e., same spatial regions) among pairs of different concepts. This eventually l
eads to the model being unable to distinguish between the two concepts and one o
f them being ignored in the final generation. Next, while these models attempt t
o capture all such concepts during the beginning of denoising (e.g., first few s
teps) as evidenced by cross-attention maps, this knowledge is not retained by th
e end of denoising (e.g., last few steps). Such loss of knowledge eventually lea
ds to inaccurate generation outputs. To address these issues, our key innovation
s include two test-time attention-based loss functions that substantially improv
e the performance of pretrained baseline text-to-image diffusion models. First,
our attention segregation loss reduces the cross-attention overlap between atten
tion maps of different concepts in the text prompt, thereby reducing the confusi
on/conflict among various concepts and the eventual capture of all concepts in t
he generated output. Next, our attention retention loss explicitly forces text-t
o-image diffusion models to retain cross-attention information for all concepts
across all denoising time steps, thereby leading to reduced information loss and
 the preservation of all concepts in the generated output. We conduct extensive
experiments with the proposed loss functions on a variety of text prompts and de
monstrate they lead to generated images that are significantly semantically clos
er to the input text when compared to baseline text-to-image diffusion models.
****

LivePose: Online 3D Reconstruction from Monocular Video with Dynamic Camera Pose
s
Noah Stier, Baptiste Angles, Liang Yang, Yajie Yan, Alex Colburn, Ming Chuang; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 7921-7930
Dense 3D reconstruction from RGB images traditionally assumes static camera pose
 estimates. This assumption has endured, even as recent works have increasingly
focused on real-time methods for mobile devices. However, the assumption of a fi
xed pose for each image does not hold for online execution: poses from real-time
 SLAM are dynamic and may be updated following events such as bundle adjustment
and loop closure. This has been addressed in the RGB-D setting, by de-integratin
g past views and re-integrating them with updated poses, but it remains largely
untreated in the RGB-only setting. We formalize this problem to define the new t
ask of dense online reconstruction from dynamically-posed images. To support fur
ther research, we introduce a dataset called LivePose containing the dynamic pos
es from a SLAM system running on ScanNet. We select three recent reconstruction
systems and apply a framework based on de-integration to adapt each one to the d
ynamic-pose setting. In addition, we propose a novel, non-linear de-integration
module that learns to remove stale scene content. We show that responding to pos
e updates is critical for high-quality reconstruction, and that our de-integrati
on framework is an effective solution.
****

Efficient Joint Optimization of Layer-Adaptive Weight Pruning in Deep Neural Net
works
Kaixin Xu, Zhe Wang, Xue Geng, Min Wu, Xiaoli Li, Weisi Lin; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17447-174
57

In this paper, we propose a novel layer-adaptive weight-pruning approach for Dee p Neural Networks (DNNs) that addresses the challenge of optimizing the output d istortion minimization while adhering to a target pruning ratio constraint. Our approach takes into account the collective influence of all layers to design a l ayer-adaptive pruning scheme. We discover and utilize a very important additivit y property of output distortion caused by pruning weights on multiple layers. Th is property enables us to formulate the pruning as a combinatorial optimization problem and efficiently solve it through dynamic programming. By decomposing the problem into sub-problems, we achieve linear time complexity, making our optimi zation algorithm fast and feasible to run on CPUs. Our extensive experiments dem onstrate the superiority of our approach over existing methods on the ImageNet a nd CIFAR-10 datasets. On CIFAR-10, our method achieves remarkable improvements, outperforming others by up to 1.0% for ResNet-32, 0.5% for VGG-16, and 0.7% for DenseNet-121 in terms of top-1 accuracy. On ImageNet, we achieve up to 4.7% and 4.6% higher top-1 accuracy compared to other methods for VGG-16 and ResNet-50, r espectively. These results highlight the effectiveness and practicality of our a pproach for enhancing DNN performance through layer-adaptive weight pruning. Code will be available on https://github.com/Akimoto-Cris/RD_VIT_PRUNE.
****
Feature Modulation Transformer: Cross-Refinement of Global Representation via Hi gh-Frequency Prior for Image Super-Resolution
Ao Li, Le Zhang, Yun Liu, Ce Zhu; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2023, pp. 12514-12524
Transformer-based methods have exhibited remarkable potential in single image su per-resolution (SISR) by effectively extracting long-range dependencies. However , most of the current research in this area has prioritized the design of transf ormer blocks to capture global information, while overlooking the importance of incorporating high-frequency priors, which we believe could be beneficial. In ou r study, we conducted a series of experiments and found that transformer structu res are more adept at capturing low-frequency information, but have limited capa city in constructing high-frequency representations when compared to their convo lutional counterparts. Our proposed solution, the cross-refinement adaptive feat ure modulation transformer (CRAFT), integrates the strengths of both convolution al and transformer structures. It comprises three key components: the high-frequ ency enhancement residual block (HFERB) for extracting high-frequency informatio n, the shift rectangle window attention block (SRWAB) for capturing global infor mation, and the hybrid fusion block (HFB) for refining the global representation . Our experiments on multiple datasets demonstrate that CRAFT outperforms state-of-the-art methods by up to 0.29dB while using fewer parameters. The source code will be made available at: https://github.com/AVC2-UESTC/CRAFT-SR.git.
****
Exploring the Sim2Real Gap Using Digital Twins
Sruthi Sudhakar, Jon Hanzelka, Josh Bobillot, Tanmay Randhavane, Neel Joshi, Vib hav Vineet; Proceedings of the IEEE/CVF International Conference on Computer Vis ion (ICCV), 2023, pp. 20418-20427
It is very time consuming to create datasets for training computer vision models . An emerging alternative is to use synthetic data, but if the synthetic data is not similar enough to the real data, the performance is typically below that of training with real data. Thus using synthetic data still requires a large amoun t of time, money, and skill as one needs to author the data carefully. In this p aper, we seek to understand which aspects of this authoring process are most cri tical. We present an analysis of which factors of variation between simulated an d real data are most important. We capture images of YCB objects to create a nov el YCB-Real dataset. We then create a novel synthetic "digital twin" dataset, YC B-Synthetic, which matches the YCB-Real dataset and includes variety of artifact s added to the synthetic data. We study the affects of these artifacts on our da taset and two existing published datasets on two different computer vision tasks : object detection and instance segmentation. We provide an analysis of the cost -benefit trade-offs between artist time for fixing artifacts and trained model a ccuracy. We plan to release this dataset (images and 3D assets) so they can be f

urther used by the community.
****
MPI-Flow: Learning Realistic Optical Flow with Multiplane Images
Yingping Liang, Jiaming Liu, Debing Zhang, Ying Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13857-13868
The accuracy of learning-based optical flow estimation models heavily relies on the realism of the training datasets. Current approaches for generating such datasets either employ synthetic data or generate images with limited realism. However, the domain gap of these data with real-world scenes constrains the generalization of the trained model to real-world applications. To address this issue, we investigate generating realistic optical flow datasets from real-world images. Firstly, to generate highly realistic new images, we construct a layered depth representation, known as multiplane images (MPI), from single-view images. This allows us to generate novel view images that are highly realistic. To generate optical flow maps that correspond accurately to the new image, we calculate the optical flows of each plane using the camera matrix and plane depths. We then project these layered optical flows into the output optical flow map with volume rendering. Secondly, to ensure the realism of motion, we present an independent object motion module that can separate the camera and dynamic object motion in MPI. This module addresses the deficiency in MPI-based single-view methods, where optical flow is generated only by camera motion and does not account for any object movement. We additionally devise a depth-aware inpainting module to merge new images with dynamic objects and address unnatural motion occlusions. We show the superior performance of our method through extensive experiments on real-world datasets. Moreover, our approach achieves state-of-the-art performance in both unsupervised and supervised training of learning-based models. The code will be made publicly available at: https://github.com/Sharpiless/MPI-Flow.
****
Re:PolyWorld - A Graph Neural Network for Polygonal Scene Parsing
Stefano Zorzi, Friedrich Fraundorfer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16762-16771
While most state-of-the-art instance segmentation methods produce pixel-wise segmentation masks, numerous applications demand precise vector polygons of detected objects instead of rasterized output. This paper proposes Re:PolyWorld as a remastered and improved version of PolyWorld, a neural network that extracts object vertices from an image and connects them optimally to generate precise polygons. The objective of this work was to overcome weaknesses and shortcomings of the original model, as well as introducing an improved polygonal representation to obtain a general-purpose method for polygon extraction in images. The architecture has been redesigned to not only exploit vertex features, but to also make use of the visual appearance of edges. To this end, an edge-aware Graph Neural Network predicts the connection strength between each pair of vertices, which is further used to compute the assignment by solving a differentiable optimal transport problem. The proposed redefinition of the polygonal scene turns the method into a powerful generalized approach that can be applied to a large variety of tasks and problem settings, such as building extraction, floorplan reconstruction and even wireframe parsing. Re:PolyWorld not only outperforms the original model on building extraction in aerial images, thanks to the proposed joint analysis of vertices and edges, but also beats the state-of-the-art in multiple other domains.
****
FaceCLIPNeRF: Text-driven 3D Face Manipulation using Deformable Neural Radiance Fields
Sungwon Hwang, Junha Hyung, Daejin Kim, Min-Jung Kim, Jaegul Choo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3469-3479
As recent advances in Neural Radiance Fields (NeRF) have enabled high-fidelity 3D face reconstruction and novel view synthesis, its manipulation also became an essential task in 3D vision. However, existing manipulation methods require extensive human labor, such as a user-provided semantic mask and manual attribute se

arch unsuitable for non-expert users. Instead, our approach is designed to requi
re a single text to manipulate a face reconstructed with NeRF. To do so, we firs
t train a scene manipulator, a latent code-conditional deformable NeRF, over a d
ynamic scene to control a face deformation using the latent code. However, repre
senting a scene deformation with a single latent code is unfavorable for composi
ting local deformations observed in different instances. As so, our proposed Pos
ition-conditional Anchor Compositor (PAC) learns to represent a manipulated scen
e with spatially varying latent codes. Their renderings with the scene manipulat
or are then optimized to yield high cosine similarity to a target text in CLIP e
mbedding space for text-driven manipulation. To the best of our knowledge, our a
pproach is the first to address the text-driven manipulation of a face reconstru
cted with NeRF. Extensive results, comparisons, and ablation studies demonstrate
 the effectiveness of our approach.
****
Video State-Changing Object Segmentation
Jiangwei Yu, Xiang Li, Xinran Zhao, Hongming Zhang, Yu-Xiong Wang; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 204
39-20448
Daily objects commonly experience state changes. For example, slicing a cucumber
 changes its state from whole to sliced. Learning about object state changes in
Video Object Segmentation (VOS) is crucial for understanding and interacting wit
h the visual world. Conventional VOS benchmarks do not consider this challenging
 yet crucial problem. This paper makes a pioneering effort to introduce a weakly
-supervised benchmark on Video State-Changing Object Segmentation (VSCOS). We co
nstruct our VSCOS benchmark by selecting state-changing videos from existing dat
asets. In advocate of an annotation-efficient approach towards state-changing ob
ject segmentation, we only annotate the first and last frames of training videos
, which is different from conventional VOS. Notably, an open-vocabulary setting
is included to evaluate the generalization to novel types of objects or state ch
anges. We empirically illustrate that state-of-the-art VOS models struggle with
state-changing objects and lose track after the state changes. We analyze the ma
in difficulties of our VSCOS task and identify three technical improvements, nam
ely, fine-tuning strategies, representation learning, and integrating motion inf
ormation. Applying these improvements results in a strong baseline for segmentin
g state-changing objects consistently. Our benchmark and baseline methods are pu
blicly available at https://github.com/venom12138/VSCOS.
****
Learning Shape Primitives via Implicit Convexity Regularization
Xiaoyang Huang, Yi Zhang, Kai Chen, Teng Li, Wenjun Zhang, Bingbing Ni; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 3642-3651
Shape primitives decomposition has been an important and long-standing task in 3
D shape analysis. Prior arts heavily rely on 3D point clouds or voxel data for s
hape primitives extraction, which are less practical in real-world scenarios. Th
is paper proposes to learn shape primitives from multi-view images by introducin
g implicit surface rendering. It is challenging since implicit shapes have a hig
h degree of freedom, which violates the simplicity property of shape primitives.
 In this work, a novel regularization term named Implicit Convexity Regularizati
on (ICR) imposed on implicit primitive learning is proposed to tackle this probl
em. We start with the convexity definition of general 3D shapes, and then derive
 the equivalent expression for implicit shapes represented by signed distance fu
nctions (SDFs). Further, instead of directly constraining the output SDF values
which cause unstable optimization, we alternatively impose constraint on second
order directional derivatives on line segments inside the shapes, which proves t
o be a tighter condition for 3D convexity. Implicit primitives constrained by th
e proposed ICR are combined into a whole object via softmax-weighted-sum operati
on over all primitive SDFs. Experiments on synthetic and real-world datasets sho
w that our method is able to decompose objects into simple and reasonable shape
primitives without the need of segmentation labels or 3D data. Code and data is
publicly available in https://github.com/seanywang0408/ICR.

****
MonoNeRF: Learning a Generalizable Dynamic Radiance Field from Monocular Videos
Fengrui Tian, Shaoyi Du, Yueqi Duan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17903-17913
In this paper, we target at the problem of learning a generalizable dynamic radiance field from monocular videos. Different from most existing NeRF methods that are based on multiple views, monocular videos only contain one view at each timestamp, thereby suffering from ambiguity along the view direction in estimating point features and scene flows. Previous studies such as DynNeRF disambiguate point features by positional encoding, which is not transferable and severely limits the generalization ability. As a result, these methods have to train one independent model for each scene and suffer from heavy computational costs when applying to increasing monocular videos in real-world applications. To address this, We propose MonoNeRF to simultaneously learn point features and scene flows with point trajectory and feature correspondence constraints across frames. More specifically, we learn an implicit velocity field to estimate point trajectory from temporal features with Neural ODE, which is followed by a flow-based feature aggregation module to obtain spatial features along the point trajectory. We jointly optimize temporal and spatial features in an end-to-end manner. Experiments show that our MonoNeRF is able to learn from multiple scenes and support new applications such as scene editing, unseen frame synthesis, and fast novel scene adaptation. Codes are available at https://github.com/tianfr/MonoNeRF
****
PG-RCNN: Semantic Surface Point Generation for 3D Object Detection
Inyong Koo, Inyoung Lee, Se-Ho Kim, Hee-Seon Kim, Woo-jin Jeon, Changick Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18142-18151
One of the main challenges in LiDAR-based 3D object detection is that the sensors often fail to capture the complete spatial information about the objects due to long distance and occlusion.
Two-stage detectors with point cloud completion approaches tackle this problem by adding more points to the regions of interest (RoIs) with a pre-trained network.
However, these methods generate dense point clouds of objects for all region proposals, assuming that objects always exist in the RoIs. This leads to the indiscriminate point generation for incorrect proposals as well.
Motivated by this, we propose Point Generation R-CNN (PG-RCNN), a novel end-to-end detector that generates semantic surface points of foreground objects for accurate detection.
Our method uses a jointly trained RoI point generation module to process the contextual information of RoIs and estimate the complete shape and displacement of foreground objects.
For every generated point, PG-RCNN assigns a semantic feature that indicates the estimated foreground probability.
Extensive experiments show that the point clouds generated by our method provide geometrically and semantically rich information for refining false positive and misaligned proposals.
PG-RCNN achieves competitive performance on the KITTI benchmark, with significantly fewer parameters than state-of-the-art models.
The code is available at https://github.com/quotation2520/PG-RCNN.
****
ITI-GEN: Inclusive Text-to-Image Generation
Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, Fernando De la Torre; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3969-3980
Text-to-image generative models often reflect the biases of the training data, leading to unequal representations of underrepresented groups. This study investigates inclusive text-to-image generative models that generate images based on human-written prompts and ensure the resulting images are uniformly distributed across attributes of interest. Unfortunately, directly expressing the desired attr

ibutes in the prompt often leads to sub-optimal results due to linguistic ambiguity or model misrepresentation. Hence, this paper proposes a drastically different approach that adheres to the maxim that "a picture is worth a thousand words". We show that, for some attributes, images can represent concepts more expressively than text. For instance, categories of skin tones are typically hard to specify by text but can be easily represented by example images. Building upon these insights, we propose a novel approach, ITI-GEN, that leverages readily available reference images for Inclusive Text-to-Image GENeration. The key idea is learning a set of prompt embeddings to generate images that can effectively represent all desired attribute categories. More importantly, ITI-GEN requires no model fine-tuning, making it computationally efficient to augment existing text-to-image models. Extensive experiments demonstrate that ITI-GEN largely improves over state-of-the-art models to generate inclusive images from a prompt.
****

Learning Depth Estimation for Transparent and Mirror Surfaces

Alex Costanzino, Pierluigi Zama Ramirez, Matteo Poggi, Fabio Tosi, Stefano Mattoccia, Luigi Di Stefano; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9244-9255

Inferring the depth of transparent or mirror (ToM) surfaces represents a hard challenge for either sensors, algorithms, or deep networks. We propose a simple pipeline for learning to estimate depth properly for such surfaces with neural networks, without requiring any ground-truth annotation. We unveil how to obtain reliable pseudo labels by in-painting ToM objects in images and processing them with a monocular depth estimation model. These labels can be used to fine-tune existing monocular or stereo networks, to let them learn how to deal with ToM surfaces. Experimental results on the Booster dataset show the dramatic improvements enabled by our remarkably simple proposal.
****

Learning Neural Eigenfunctions for Unsupervised Semantic Segmentation

Zhijie Deng, Yucen Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 551-561

Unsupervised semantic segmentation is a long-standing challenge in computer vision with great significance. Spectral clustering is a theoretically grounded solution to it where the spectral embeddings for pixels are computed to construct distinct clusters. Despite recent progress in enhancing spectral clustering with powerful pre-trained models, current approaches still suffer from inefficiencies in spectral decomposition and inflexibility in applying them to the test data. This work addresses these issues by casting spectral clustering as a parametric approach that employs neural network-based eigenfunctions to produce spectral embeddings. The outputs of the neural eigenfunctions are further restricted to discrete vectors that indicate clustering assignments directly. As a result, an end-to-end NN-based paradigm of spectral clustering emerges. In practice, the neural eigenfunctions are lightweight and take the features from pre-trained models as inputs, improving training efficiency and unleashing the potential of pre-trained models for dense prediction. We conduct extensive empirical studies to validate the effectiveness of our approach and observe significant performance gains over competitive baselines on Pascal Context, Cityscapes, and ADE20K benchmarks. The code is available at https://github.com/thudzj/NeuralEigenfunctionSegmentor.
****

Shape Analysis of Euclidean Curves under Frenet-Serret Framework

Perrine Chassat, Juhyun Park, Nicolas Brunel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4027-4036

Geometric frameworks for analyzing curves are common in applications as they focus on invariant features and provide visually satisfying solutions to standard problems such as computing invariant distances, averaging curves, or registering curves. We show that for any smooth curve in R^d, d>1, the generalized curvatures associated with the Frenet-Serret equation can be used to define a Riemannian geometry that takes into account all the geometric features of the shape. This geometry is based on a Square Root Curvature Transform that extends the square root-velocity transform for Euclidean curves (in any dimensions) and provides like

ly geodesics that avoid artefacts encountered by representations using only first-order geometric information. Our analysis is supported by simulated data and is especially relevant for analyzing human motions. We consider trajectories acquired from sign language, and show the interest of considering curvature and also torsion in their analysis, both being physically meaningful.
****
Representation Uncertainty in Self-Supervised Learning as Variational Inference

Hiroki Nakamura, Masashi Okada, Tadahiro Taniguchi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16484-16493

In this study, a novel self-supervised learning (SSL) method is proposed, which considers SSL in terms of variational inference to learn not only representation but also representation uncertainties. SSL is a method of learning representations without labels by maximizing the similarity between image representations of different augmented views of an image. Meanwhile, variational autoencoder (VAE) is an unsupervised representation learning method that trains a probabilistic generative model with variational inference. Both VAE and SSL can learn representations without labels, but their relationship has not been investigated in the past. Herein, the theoretical relationship between SSL and variational inference has been clarified. Furthermore, a novel method, namely variational inference SimSiam (VI-SimSiam), has been proposed. VI-SimSiam can predict the representation uncertainty by interpreting SimSiam with variational inference and defining the latent space distribution. The present experiments qualitatively show that VI-SimSiam could learn uncertainty by comparing input images and predicted uncertainties. Additionally, we described a relationship between estimated uncertainty and classification accuracy.
****
Efficient Diffusion Training via Min-SNR Weighting Strategy

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, Baining Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7441-7451

Denoising diffusion models have been a mainstream approach for image generation, however, training these models often suffers from slow convergence. In this paper, we discovered that the slow convergence is partly due to conflicting optimization directions between timesteps. To address this issue, we treat the diffusion training as a multi-task learning problem, and introduce a simple yet effective approach referred to as Min-SNR-g. This method adapts loss weights of timesteps based on clamped signal-to-noise ratios, which effectively balances the conflicts among timesteps. Our results demonstrate a significant improvement in converging speed, 3.4x faster than previous weighting strategies. It is also more effective, achieving a new record FID score of 2.06 on the ImageNet 256x256 benchmark using smaller architectures than that employed in previous state-of-the-art.
****
Bridging Vision and Language Encoders: Parameter-Efficient Tuning for Referring Image Segmentation

Zunnan Xu, Zhihong Chen, Yong Zhang, Yibing Song, Xiang Wan, Guanbin Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17503-17512

Parameter efficient tuning (PET) has received considerable attention owing to its applicability to reduce the number of parameters that need to be updated while maintaining competitive performance and providing better hardware resource savings. Although substantial progress has been made, most existing studies mainly focus on either single-modal tasks or simple classification tasks, with few works paying attention to the dense prediction tasks and the interaction between different modalities. Therefore, in this paper, we do an in-depth investigation of the efficient tuning problem on referring image segmentation. First, considering the absence of interaction between the dual encoder, we design a novel adapter named Bridger to facilitate the exchange of cross-modal information. This module also plays a role in injecting vision-specific inductive biases and task-specific information into the pre-trained model while keeping its original parameters fixed. Second, we design a lightweight decoder for referring image segmentation t

o make further alignment on visual and linguistic features. To perform a compreh
ensive assessment and promote further research, we evaluate the proposed framewo
rk on several challenging benchmarks. Experimental results illustrate the effect
iveness of our approach. Updating only 1.61% to 3.38% parameters, the proposed f
ramework gains comparable or even superior performance compared to existing full
 fine-tuning methods that utilize the same backbone.
****
Towards Zero-Shot Scale-Aware Monocular Depth Estimation
Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rare■ Ambru■, Adrien Gaidon; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 9233-9243
Monocular depth estimation is scale-ambiguous, and thus requires scale supervisi
on to produce metric predictions. Even so, the resulting models will be geometry
-specific, with learned scales that cannot be directly transferred across domain
s. Because of that, recent works focus instead on relative depth, eschewing scal
e in favor of improved up-to-scale zero-shot transfer. In this work we introduce
 ZeroDepth, a novel monocular depth estimation framework capable of predicting m
etric scale for arbitrary test images from different domains and camera paramete
rs. This is achieved by (i) the use of input-level geometric embeddings that ena
ble the network to learn a scale prior over objects; and (ii) decoupling the enc
oder and decoder stages, via a variational latent representation that is conditi
oned on single frame information. We evaluated ZeroDepth targeting both outdoor
(KITTI, DDAD, nuScenes) and indoor (NYUv2) benchmarks, and achieved a new state-
of-the-art in both settings using the same pre-trained model, outperforming meth
ods that train on in-domain data and require test-time scaling to produce metric
 estimates.
****
ATT3D: Amortized Text-to-3D Object Synthesis
Jonathan Lorraine, Kevin Xie, Xiaohui Zeng, Chen-Hsuan Lin, Towaki Takikawa, Nic
holas Sharp, Tsung-Yi Lin, Ming-Yu Liu, Sanja Fidler, James Lucas; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 179
46-17956
Text-to-3D modelling has seen exciting progress by combining generative text-to-
image models with image-to-3D methods like Neural Radiance Fields. DreamFusion r
ecently achieved high-quality results but requires a lengthy, per-prompt optimiz
ation to create 3D objects. To address this, we amortize optimization over text
prompts by training on many prompts simultaneously with a unified model instead
of separately. With this, we share computation across a prompt set, training in
less time than per-prompt optimization. Our framework, Amortized Text-to-3D (ATT
3D), enables knowledge sharing between prompts to generalize to unseen setups an
d smooth interpolations between text for novel assets and simple animations.
****
Virtual Try-On with Pose-Garment Keypoints Guided Inpainting
Zhi Li, Pengfei Wei, Xiang Yin, Zejun Ma, Alex C. Kot; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2023, pp. 22788-22797
Virtual try-on is an important technology supporting online apparel shopping, wh
ich provides consumers with a virtual experience to fit garments without physica
lly wearing them. Recently, the image-based virtual try-on has received growing
research attention. However, the synthetic results of existing virtual try-on me
thods usually present distortions in garment shape and lose pattern details. In
this paper, we propose a pose-garment keypoints guided inpainting method for the
 image-based virtual try-on task, which produces high-fidelity try-on images and
 well preserves the shapes and patterns of the garments. In our method, human po
se and garment keypoints are extracted from source images and constructed as gra
phs to predict the garment keypoints at the target pose. After which, the predic
ted keypoints are used as guide information to predict the target segmentation m
ap and warp the garment image. The try-on image is finally generated with a sema
ntic-conditioned inpainting scheme using the segmentation map and recomposed per
son image as conditions. To verify the effectiveness of our proposed method, we
conduct extensive experiments on the VITON-HD dataset under both paired and unpa

ired experimental settings. The qualitative and quantitative results show that our method significantly outperforms prior methods at different image resolutions. The codes repository link is https://github.com/lizhi-ntu/KGI.
****

Learning by Sorting: Self-supervised Learning with Group Ordering Constraints
Nina Shvetsova, Felix Petersen, Anna Kukleva, Bernt Schiele, Hilde Kuehne; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16453-16463
Contrastive learning has become an important tool in learning representations from unlabeled data mainly relying on the idea of minimizing distance between positive data pairs, e.g., views from the same images, and maximizing distance between negative data pairs, e.g., views from different images. This paper proposes a new variation of the contrastive learning objective, Group Ordering Constraints (GroCo), that leverages the idea of sorting the distances of positive and negative pairs and computing the respective loss based on how many positive pairs have a larger distance than the negative pairs, and thus are not ordered correctly. To this end, the GroCo loss is based on differentiable sorting networks, which enable training with sorting supervision by matching a differentiable permutation matrix, which is produced by sorting a given set of scores, to a respective ground truth permutation matrix. Applying this idea to groupwise pre-ordered inputs of multiple positive and negative pairs allows introducing the GroCo loss with implicit emphasis on strong positives and negatives, leading to better optimization of the local neighborhood. We evaluate the proposed formulation on various self-supervised learning benchmarks and show that it not only leads to improved results compared to vanilla contrastive learning but also shows competitive performance to comparable methods in linear probing and outperforms current methods in k-NN performance.
****

Cross Modal Transformer: Towards Fast and Robust 3D Object Detection
Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, Xiangyu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18268-18278
In this paper, we propose a robust 3D detector, named Cross Modal Transformer (CMT), for end-to-end 3D multi-modal detection. Without explicit view transformation, CMT takes the image and point clouds tokens as inputs and directly outputs accurate 3D bounding boxes. The spatial alignment of multi-modal tokens is performed by encoding the 3D points into multi-modal features. The core design of CMT is quite simple while its performance is impressive. It achieves 74.1% NDS (state-of-the-art with single model) on nuScenes test set while maintaining faster inference speed. Moreover, CMT has a strong robustness even if the LiDAR is missing. Code is released at https: //github.com/junjie18/CMT .
****

Perceptual Grouping in Contrastive Vision-Language Models
Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, Jonathon Shlens; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5571-5584
Recent advances in zero-shot image recognition suggest that vision-language models learn generic visual representations with a high degree of semantic information that may be arbitrarily probed with natural language phrases. Understanding an image, however, is not just about understanding what content resides within an image, but importantly, where that content resides. In this work we examine how well vision-language models are able to understand where objects reside within an image and group together visually related parts of the imagery. We demonstrate how contemporary vision and language representation learning models based on contrastive losses and large web-based data capture limited object localization information. We propose a minimal set of modifications that results in models that uniquely learn both semantic and spatial information. We measure this performance in terms of zero-shot image recognition, unsupervised bottom-up and top-down semantic segmentations, as well as robustness analyses. We find that the resulting model achieves state-of-the-art results in terms of unsupervised segmentatio

n, and demonstrate that the learned representations are uniquely robust to spuri ous correlations in datasets designed to probe the causal behavior of vision mod els.
****

Dynamic Perceiver for Efficient Visual Recognition
Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, Gao Huang; Proceedings of the IEEE/CVF International Co nference on Computer Vision (ICCV), 2023, pp. 5992-6002
Early exiting has become a promising approach to im- proving the inference effic iency of deep networks. By structuring models with multiple classifiers (exits), predictions for "easy" samples can be generated at earlier exits, negating the need for executing deeper layers. Current multi-exit networks typically implemen t linear classifiers at intermediate layers, compelling low-level features to en capsulate high-level semantics. This sub-optimal design invariably undermines th e performance of later exits. In this paper, we propose Dynamic Perceiver (Dyn-P erceiver) to decouple the feature extraction procedure and the early classificat ion task with a novel dual-branch architecture. A feature branch serves to extra ct image features, while a classification branch processes a latent code assigne d for classification tasks. Bi-directional cross-attention layers are establishe d to progressively fuse the information of both branches. Early exits are placed exclusively within the classification branch, thus eliminating the need for lin ear separability in low-level features. Dyn-Perceiver constitutes a versatile an d adaptable framework that can be built upon various architectures. Experiments on image classification, action recognition, and object detection demonstrate th at our method significantly improves the inference efficiency of different backb ones, outperforming numerous competitive approaches across a broad range of comp utational budgets. Evaluation on both CPU and GPU platforms substantiate the sup erior practical efficiency of Dyn-Perceiver. Code is available at https://www.gi thub. com/LeapLabTHU/Dynamic_Perceiver.
****

MoTIF: Learning Motion Trajectories with Local Implicit Neural Functions for Con tinuous Space-Time Video Super-Resolution
Yi-Hsin Chen, Si-Cun Chen, Yi-Hsin Chen, Yen-Yu Lin, Wen-Hsiao Peng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2 3131-23141
This work addresses continuous space-time video super-resolution (C-STVSR) that aims to up-scale an input video both spatially and temporally by any scaling fac tors. One key challenge of C-STVSR is to propagate information temporally among the input video frames. To this end, we introduce a space-time local implicit ne ural function. It has the striking feature of learning forward motion for a cont inuum of pixels. We motivate the use of forward motion from the perspective of l earning individual motion trajectories, as opposed to learning a mixture of moti on trajectories with backward motion. To ease motion interpolation, we encode sp arsely sampled forward motion extracted from the input video as the contextual i nput. Along with a reliability-aware splatting and decoding scheme, our framewor k, termed MoTIF, achieves the state-of-the-art performance on C-STVSR. The sourc e code of MoTIF is available at https://github.com/sichun233746/MoTIF.
****

CRN: Camera Radar Net for Accurate, Robust, Efficient 3D Perception
Youngseok Kim, Juyeb Shin, Sanmin Kim, In-Jae Lee, Jun Won Choi, Dongsuk Kum; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 23, pp. 17615-17626
Autonomous driving requires an accurate and fast 3D perception system that inclu des 3D object detection, tracking, and segmentation. Although recent low-cost ca mera-based approaches have shown promising results, they are susceptible to poor illumination or bad weather conditions and have a large localization error. Hen ce, fusing camera with low-cost radar, which provides precise long-range measure ment and operates reliably in all environments, is promising but has not yet bee n thoroughly investigated. In this paper, we propose Camera Radar Net (CRN), a n ovel camera-radar fusion framework that generates a semantically rich and spatia

lly accurate bird's-eye-view (BEV) feature map for various tasks. To overcome the lack of spatial information in an image, we transform perspective view image features to BEV with the help of sparse but accurate radar points. We further aggregate image and radar feature maps in BEV using multi-modal deformable attention designed to tackle the spatial misalignment between inputs. CRN with real-time setting operates at 20 FPS while achieving comparable performance to LiDAR detectors on nuScenes, and even outperforms at a far distance on 100m setting. Moreover, CRN with offline setting yields 62.4% NDS, 57.5% mAP on nuScenes test set and ranks first among all camera and camera-radar 3D object detectors.
****
PromptStyler: Prompt-driven Style Generation for Source-free Domain Generalization
Junhyeong Cho, Gilhyun Nam, Sungyeon Kim, Hunmin Yang, Suha Kwak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15702-15712
In a joint vision-language space, a text feature (e.g., from "a photo of a dog") could effectively represent its relevant image features (e.g., from dog photos). Also, a recent study has demonstrated the cross-modal transferability phenomenon of this joint space. From these observations, we propose PromptStyler which simulates various distribution shifts in the joint space by synthesizing diverse styles via prompts without using any images to deal with source-free domain generalization. The proposed method learns to generate a variety of style features (from "a S* style of a") via learnable style word vectors for pseudo-words S*. To ensure that learned styles do not distort content information, we force style-content features (from "a S* style of a [class]") to be located nearby their corresponding content features (from "[class]") in the joint vision-language space. After learning style word vectors, we train a linear classifier using synthesized style-content features. PromptStyler achieves the state of the art on PACS, VLCS, OfficeHome and DomainNet, even though it does not require any images for training.
****
Phasic Content Fusing Diffusion Model with Directional Distribution Consistency for Few-Shot Model Adaption
Teng Hu, Jiangning Zhang, Liang Liu, Ran Yi, Siqi Kou, Haokun Zhu, Xu Chen, Yabiao Wang, Chengjie Wang, Lizhuang Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2406-2415
Training a generative model with limited number of samples is a challenging task. Current methods primarily rely on few-shot model adaption to train the network. However, in scenarios where data is extremely limited (less than 10), the generative network tends to overfit and suffers from content degradation. To address these problems, we propose a novel phasic content fusing few-shot diffusion model with directional distribution consistency loss, which targets different learning objectives at distinct training stages of the diffusion model. Specifically, we design a phasic training strategy with phasic content fusion to help our model learn content and style information when t is large, and learn local details of target domain when t is small, leading to an improvement in the capture of content, style and local details. Furthermore, we introduce a novel directional distribution consistency loss that ensures the consistency between the generated and source distributions more efficiently and stably than the prior methods, preventing our model from overfitting. Finally, we propose a cross-domain structure guidance strategy that enhances structure consistency during domain adaptation. Theoretical analysis, qualitative and quantitative experiments demonstrate the superiority of our approach in few-shot generative model adaption tasks compared to state-of-the-art methods.
****
SVQNet: Sparse Voxel-Adjacent Query Network for 4D Spatio-Temporal LiDAR Semantic Segmentation
Xuechao Chen, Shuangjie Xu, Xiaoyi Zou, Tongyi Cao, Dit-Yan Yeung, Lu Fang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8569-8578

LiDAR-based semantic perception tasks are critical yet challenging for autonomous driving. Due to the motion of objects and static/dynamic occlusion, temporal information plays an essential role in reinforcing perception by enhancing and completing single-frame knowledge. Previous approaches either directly stack historical frames to the current frame or build a 4D spatio-temporal neighborhood using KNN, which duplicates computation and hinders real-time performance. Based on our observation that stacking all the historical points would damage performance due to a large amount of redundant and misleading information, we propose the Sparse Voxel-Adjacent Query Network (SVQNet) for 4D LiDAR semantic segmentation. To take full advantage of the historical frames high-efficiently, we shunt the historical points into two groups with reference to the current points. One is the Voxel-Adjacent Neighborhood carrying local enhancing knowledge. The other is the Historical Context completing the global knowledge. Then we propose new modules to select and extract the instructive features from the two groups. Our SVQNet achieves state-of-the-art performance in LiDAR semantic segmentation of the SemanticKITTI benchmark and the nuScenes dataset.
****
HAL3D: Hierarchical Active Learning for Fine-Grained 3D Part Labeling
Fenggen Yu, Yiming Qian, Francisca Gil-Ureta, Brian Jackson, Eric Bennett, Hao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 865-875
We present the first active learning tool for fine-grained 3D part labeling, a problem which challenges even the most advanced deep learning (DL) methods due to the significant structural variations among the intricate parts. For the same reason, the necessary effort to annotate training data is tremendous, motivating approaches to minimize human involvement. Our labeling tool iteratively verifies or modifies part labels predicted by a deep neural network, with human feedback continually improving the network prediction. To effectively reduce human efforts, we develop two novel features in our tool, hierarchical and symmetry-aware active labeling. Our human-in-the-loop approach, coined HAL3D, achieves close to error-free fine-grained annotations on any test set with pre-defined hierarchical part labels, with 80% time-saving over manual effort. We will release the finely labeled models to serve the community.
****
MEFLUT: Unsupervised 1D Lookup Tables for Multi-exposure Image Fusion
Ting Jiang, Chuan Wang, Xinpeng Li, Ru Li, Haoqiang Fan, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10542-10551
In this paper, we introduce a new approach for high-quality multi-exposure image fusion (MEF). We show that the fusion weights of an exposure can be encoded into a 1D lookup table (LUT), which takes pixel intensity value as input and produces fusion weight as output. We learn one 1D LUT for each exposure, then all the pixels from different exposures can query 1D LUT of that exposure independently for high-quality and efficient fusion. Specifically, to learn these 1D LUTs, we involve attention mechanism in various dimensions including frame, channel and spatial ones into the MEF task so as to bring us significant quality improvement over the state-of-the-art (SOTA). In addition, we collect a new MEF dataset consisting of 960 samples, 155 of which are manually tuned by professionals as ground-truth for evaluation. Our network is trained by this dataset in an unsupervised manner. Extensive experiments are conducted to demonstrate the effectiveness of all the newly proposed components, and results show that our approach outperforms the SOTA in our and another representative dataset SICE, both qualitatively and quantitatively. Moreover, our 1D LUT approach takes less than 4ms to run a 4K image on a PC GPU. Given its high quality, efficiency and robustness, our method has been shipped into millions of Android mobiles across multiple brands world-wide. Code is available at: https://github.com/Hedlen/MEFLUT.
****
FedPerfix: Towards Partial Model Personalization of Vision Transformers in Federated Learning
Guangyu Sun, Matias Mendieta, Jun Luo, Shandong Wu, Chen Chen; Proceedings of th

e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4988-49
98
Personalized Federated Learning (PFL) represents a promising solution for decent
ralized learning in heterogeneous data environments. Partial model personalizati
on has been proposed to improve the efficiency of PFL by selectively updating lo
cal model parameters instead of aggregating all of them. However, previous work
on partial model personalization has mainly focused on Convolutional Neural Netw
orks (CNNs), leaving a gap in understanding how it can be applied to other popul
ar models such as Vision Transformers (ViTs).
In this work, we investigate where and how to partially personalize a ViT model
. Specifically, we empirically evaluate the sensitivity to data distribution of
each type of layer. Based on the insights that the self-attention layer and the
classification head are the most sensitive parts of a ViT, we propose a novel ap
proach called FedPerfix, which leverages plugins to transfer information from th
e aggregated model to the local client as a personalization. Finally, we evaluat
e the proposed approach on CIFAR-100, OrganAMNIST, and Office-Home datasets and
demonstrate its effectiveness in improving the model's performance compared to s
everal advanced PFL methods.
****
Conditional 360-degree Image Synthesis for Immersive Indoor Scene Decoration
Ka Chun Shum, Hong-Wing Pang, Binh-Son Hua, Duc Thanh Nguyen, Sai-Kit Yeung; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
3, pp. 4478-4488
In this paper, we address the problem of conditional scene decoration for 360deg
images. Our method takes a 360deg background photograph of an indoor scene and
generates decorated images of the same scene in the panorama view. To do this, w
e develop a 360-aware object layout generator that learns latent object vectors
in the 360deg view to enable a variety of furniture arrangements for an input 36
0deg background image. We use this object layout to condition a generative adver
sarial network to synthesize images of an input scene. To further reinforce the
generation capability of our model, we develop a simple yet effective scene empt
ier that removes the generated furniture and produces an emptied scene for our m
odel to learn a cyclic constraint. We train the model on the Structure3D dataset
and show that our model can generate diverse decorations with controllable obje
ct layout. Our method achieves state-of-the-art performance on the Structure3D d
ataset and generalizes well to the Zillow indoor scene dataset. Our user study c
onfirms the immersive experiences provided by the realistic image quality and fu
rniture layout in our generation results. Our implementation is available at htt
ps://github.com/kcshum/neural_360_decoration.git.
****
The Unreasonable Effectiveness of Large Language-Vision Models for Source-Free V
ideo Domain Adaptation
Giacomo Zara, Alessandro Conti, Subhankar Roy, Stéphane Lathuilière, Paolo Rota,
Elisa Ricci; Proceedings of the IEEE/CVF International Conference on Computer V
ision (ICCV), 2023, pp. 10307-10317
Source-Free Video Unsupervised Domain Adaptation (SFVUDA) task consists in adapt
ing an action recognition model, trained on a labelled source dataset, to an unl
abelled target dataset, without accessing the actual source data. The previous a
pproaches have attempted to address SFVUDA by leveraging self-supervision (e.g.,
enforcing temporal consistency) derived from the target data itself. In this wo
rk, we take an orthogonal approach by exploiting "web-supervision" from Large La
nguage-Vision Models (LLVMs), driven by the rationale that LLVMs contain a rich
world prior surprisingly robust to domain-shift. We showcase the unreasonable ef
fectiveness of integrating LLVMs for SFVUDA by devising an intuitive and paramet
er-efficient method, which we name Domain Adaptation with Large Language-Vision
models (DALL-V), that distills the world prior and complementary source model in
formation into a student network tailored for the target. Despite the simplicity
, DALL-V achieves significant improvement over state-of-the-art SFVUDA methods.
****
SIDGAN: High-Resolution Dubbed Video Generation via Shift-Invariant Learning

Urwa Muaz, Wondong Jang, Rohun Tripathi, Santhosh Mani, Wenbin Ouyang, Ravi Teja Gadde, Baris Gecer, Sergio Elizondo, Reza Madad, Naveen Nair; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7833-7842

Dubbed video generation aims to accurately synchronize mouth movements of a given facial video with driving audio while preserving identity and scene-specific visual dynamics, such as head pose and lighting. Despite the accurate lip generation of previous approaches that adopts a pretrained audio-video synchronization metric as an objective function, called Sync-Loss, extending it to high-resolution videos was challenging due to shift biases in the loss landscape that inhibit tandem optimization of Sync-Loss and visual quality, leading to a loss of detail.

To address this issue, we introduce shift-invariant learning, which generates photo-realistic high-resolution videos with accurate Lip-Sync. Further, we employ a pyramid network with coarse-to-fine image generation to improve stability and lip syncronization. Our model outperforms state-of-the-art methods on multiple benchmark datasets, including AVSpeech, HDTF, and LRW, in terms of photo-realism, identity preservation, and Lip-Sync accuracy.

****

Meta-ZSDETR: Zero-shot DETR with Meta-learning

Lu Zhang, Chenbo Zhang, Jiajia Zhao, Jihong Guan, Shuigeng Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6845-6854

Zero-shot object detection aims to localize and recognize objects of unseen classes. Most of existing works face two problems: the low recall of RPN in unseen classes and the confusion of unseen classes with background. In this paper, we present the first method that combines DETR and meta-learning to perform zero-shot object detection, named Meta-ZSDETR, where model training is formalized as an individual episode based meta-learning task. Different from Faster R-CNN based methods that firstly generate class-agnostic proposals, and then classify them with visual-semantic alignment module, Meta-ZSDETR directly predict class-specific boxes with class-specific queries and further filter them with the predicted accuracy from classification head. The model is optimized with meta-contrastive learning, which contains a regression head to generate the coordinates of class-specific boxes, a classification head to predict the accuracy of generated boxes, and a contrastive head that utilizes the proposed contrastive-reconstruction loss to further separate different classes in visual space. We conduct extensive experiments on two benchmark datasets MS COCO and PASCAL VOC. Experimental results show that our method outperforms the existing ZSD methods by a large margin.

****

GaPro: Box-Supervised 3D Point Cloud Instance Segmentation Using Gaussian Processes as Pseudo Labelers

Tuan Duc Ngo, Binh-Son Hua, Khoi Nguyen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17794-17803

Instance segmentation on 3D point clouds (3DIS) is a longstanding challenge in computer vision, where state-of-the-art methods are mainly based on full supervision. As annotating ground truth dense instance masks is tedious and expensive, solving 3DIS with weak supervision has become more practical. In this paper, we propose GaPro, a new instance segmentation for 3D point clouds using axis-aligned 3D bounding box supervision. Our two-step approach involves generating pseudo labels from box annotations and training a 3DIS network with the resulting labels. Additionally, we employ the self-training strategy to improve the performance of our method further. We devise an effective Gaussian Process to generate pseudo instance masks from the bounding boxes and resolve ambiguities when they overlap, resulting in pseudo instance masks with their uncertainty values. Our experiments show that GaPro outperforms previous weakly supervised 3D instance segmentation methods and has competitive performance compared to state-of-the-art fully supervised ones. Furthermore, we demonstrate the robustness of our approach, where we can adapt various state-of-the-art fully supervised methods to the weak s

upervision task by using our pseudo labels for training. We will release our implementation upon publication.
****
STPrivacy: Spatio-Temporal Privacy-Preserving Action Recognition
Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, Shuicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5106-5115
Existing methods of privacy-preserving action recognition (PPAR) mainly focus on frame-level (spatial) privacy removal through 2D CNNs. Unfortunately, they have two major drawbacks. First, they may compromise temporal dynamics in input videos, which are critical for accurate action recognition. Second, they are vulnerable to practical attacking scenarios where attackers probe for privacy from an entire video rather than individual frames. To address these issues, we propose a novel framework STPrivacy to perform video-level PPAR. For the first time, we introduce vision Transformers into PPAR by treating a video as a tubelet sequence, and accordingly design two complementary mechanisms, i.e., sparsification and anonymization, to remove privacy from a spatio-temporal perspective. In specific, our privacy sparsification mechanism applies adaptive token selection to abandon action-irrelevant tubelets. Then, our anonymization mechanism implicitly manipulates the remaining action-tubelets to erase privacy in the embedding space through adversarial learning. These mechanisms provide significant advantages in terms of privacy preservation for human eyes and action-privacy trade-off adjustment during deployment. We additionally contribute the first two large-scale PPAR benchmarks, VP-HMDB51 and VP-UCF101, to the community. Extensive evaluations on them, as well as two other tasks, validate the effectiveness and generalization capability of our framework.
****
Get the Best of Both Worlds: Improving Accuracy and Transferability by Grassmann Class Representation
Haoqi Wang, Zhizhong Li, Wayne Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22478-22487
We generalize the class vectors found in neural networks to linear subspaces (i.e., points in the Grassmann manifold) and show that the Grassmann Class Representation (GCR) enables simultaneous improvement in accuracy and feature transferability. In GCR, each class is a subspace, and the logit is defined as the norm of the projection of a feature onto the class subspace. We integrate Riemannian SGD into deep learning frameworks such that class subspaces in a Grassmannian are jointly optimized with the rest model parameters. Compared to the vector form, the representative capability of subspaces is more powerful. We show that on ImageNet-1K, the top-1 errors of ResNet50-D, ResNeXt50, Swin-T, and Deit3-S are reduced by 5.6%, 4.5%, 3.0%, and 3.5%, respectively. Subspaces also provide freedom for features to vary, and we observed that the intra-class feature variability grows when the subspace dimension increases. Consequently, we found the quality of GCR features is better for downstream tasks. For ResNet50-D, the average linear transfer accuracy across 6 datasets improves from 77.98% to 79.70% compared to the strong baseline of vanilla softmax. For Swin-T, it improves from 81.5% to 83.4% and for Deit3, it improves from 73.8% to 81.4%. With these encouraging results, we believe that more applications could benefit from the Grassmann class representation. Code is released at https://github.com/innerlee/GCR.
****
Computationally-Efficient Neural Image Compression with Shallow Decoders
Yibo Yang, Stephan Mandt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 530-540
Neural image compression methods have seen increasingly strong performance in recent years. However, they suffer orders of magnitude higher computational complexity compared to traditional codecs, which hinders their real-world deployment. This paper takes a step forward in closing this gap in decoding complexity by adopting shallow or even linear decoding transforms. To compensate for the resulting drop in compression performance, we exploit the often asymmetrical computation budget between encoding and decoding, by adopting more powerful encoder networor

ks and iterative encoding. We theoretically formalize the intuition behind, and our experimental results establish a new frontier in the trade-off between rate-distortion and decoding complexity for neural image compression. Specifically, w
e achieve rate-distortion performance competitive with the established mean-scal
e hyperprior architecture of Minnen et al. (2018) at less than 50K decoding FLOP
s/pixel, reducing the baseline's overall decoding complexity by 80%, or over 90%
 for the synthesis transform alone.
 Our code can be found at https://github.com/mandt-lab/shallow-ntc.
****
ObjectSDF++: Improved Object-Compositional Neural Implicit Surfaces
Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, Jianfei Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21764
-21774
In recent years, neural implicit surface reconstruction has emerged as a popular
 paradigm for multi-view 3D reconstruction. Unlike traditional multi-view stereo
 approaches, the neural implicit surface-based methods leverage neural networks
to represent 3D scenes as signed distance functions (SDFs). However, they tend t
o disregard the reconstruction of individual objects within the scene, which lim
its their performance and practical applications. To address this issue, previou
s work ObjectSDF introduced a nice framework of object-composition neural implic
it surfaces, which utilizes 2D instance masks to supervise individual object SDF
s.
  In this paper, we propose a new framework called ObjectSDF++ to overcome the l
imitations of ObjectSDF. First, in contrast to ObjectSDF whose performance is pr
imarily restricted by its converted semantic field, the core component of our mo
del is an occlusion-aware object opacity rendering formulation that directly vol
ume-renders object opacity to be supervised with instance masks. Second, we desi
gn a novel regularization term for object distinction, which can effectively mit
igate the issue that ObjectSDF may result in unexpected reconstruction in invisi
ble regions due to the lack of constraint to prevent collisions. Our extensive e
xperiments demonstrate that our novel framework not only produces superior objec
t reconstruction results but also significantly improves the quality of scene re
construction. Code and more resources can be found in https://qianyiwu.github.io
/objectsdf++
****
Tracing the Origin of Adversarial Attack for Forensic Investigation and Deterren
ce
Han Fang, Jiyi Zhang, Yupeng Qiu, Jiayang Liu, Ke Xu, Chengfang Fang, Ee-Chien C
hang; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2023, pp. 4335-4344
Deep neural networks are vulnerable to adversarial attacks. In this paper, we ta
ke the role of investigators who want to trace the attack and identify the sourc
e, that is, the particular model which the adversarial examples are generated fr
om. Techniques derived would aid forensic investigation of attack incidents and
serve as deterrence to potential attacks. We consider the buyers-seller setting
where a machine learning model is to be distributed to various buyers and each b
uyer receives a slightly different copy with same functionality. A malicious buy
er generates adversarial examples from a particular copy "Mi" and uses them to a
ttack other copies. From these adversarial examples, the investigator wants to i
dentify the source "Mi". To address this problem, we propose a two-stage separat
e-and-trace framework. The model separation stage generates multiple copies of a
 model for a same classification task. This process injects unique characteristi
cs into each copy so that adversarial examples generated have distinct and trace
able features. We give a parallel structure which pairs a unique tracer with the
 original classification model in each copy and a variational autoencoder (VAE)-
based training method to achieve this goal. The tracing stage takes in adversari
al examples and a few candidate models, and identifies the likely source. Based
on the unique features induced by the tracer, we could effectively trace the pot
ential adversarial copy by considering the output logits from each tracer. Empir
ical results show that it is possible to trace the origin of the adversarial exa

mple and the mechanism can be applied to a wide range of architectures and datasets.
****
Sketch and Text Guided Diffusion Model for Colored Point Cloud Generation
Zijie Wu, Yaonan Wang, Mingtao Feng, He Xie, Ajmal Mian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8929-8939
Diffusion probabilistic models have achieved remarkable success in text guided image generation. However, generating 3D shapes is still challenging due to the lack of sufficient data containing 3D models along with their descriptions. Moreover, text based descriptions of 3D shapes are inherently ambiguous and lack details.
 In this paper, we propose a sketch and text guided probabilistic diffusion model for colored point cloud generation that conditions the denoising process jointly with a hand drawn sketch of the object and its textual description.
 We incrementally diffuse the point coordinates and color values in a joint diffusion process to reach a Gaussian distribution. Colored point cloud generation thus amounts to learning the reverse diffusion process, conditioned by the sketch and text, to iteratively recover the desired shape and color.
 Specifically, to learn effective sketch-text embedding, our model adaptively aggregates the joint embedding of text prompt and the sketch based on a capsule attention network. Our model uses staged diffusion to generate the shape and then assign colors to different parts conditioned on the appearance prompt while preserving precise shapes from the first stage. This gives our model the flexibility to extend to multiple tasks, such as appearance re-editing and part segmentation. Experimental results demonstrate that our model outperforms the recent state-of-the-art in point cloud generation.
****
Scenimefy: Learning to Craft Anime Scene via Semi-Supervised Image-to-Image Translation
Yuxin Jiang, Liming Jiang, Shuai Yang, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7357-7367
Automatic high-quality rendering of anime scenes from complex real-world images is of significant practical value. The challenges of this task lie in the complexity of the scenes, the unique features of anime style, and the lack of high-quality datasets to bridge the domain gap. Despite promising attempts, previous efforts are still incompetent in achieving satisfactory results with consistent semantic preservation, evident stylization, and fine details. In this study, we propose Scenimefy, a novel semi-supervised image-to-image translation framework that addresses these challenges. Our approach guides the learning with structure-consistent pseudo paired data, simplifying the pure unsupervised setting. The pseudo data are derived uniquely from a semantic-constrained StyleGAN leveraging rich model priors like CLIP. We further apply segmentation-guided data selection to obtain high-quality pseudo supervision. A patch-wise contrastive style loss is introduced to improve stylization and fine details. Besides, we contribute a high-resolution anime scene dataset to facilitate future research. Our extensive experiments demonstrate the superiority of our method over state-of-the-art baselines in terms of both perceptual quality and quantitative performance.
****
Towards Unsupervised Domain Generalization for Face Anti-Spoofing
Yuchen Liu, Yabo Chen, Mengran Gou, Chun-Ting Huang, Yaoming Wang, Wenrui Dai, Hongkai Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20654-20664
Generalizable face anti-spoofing (FAS) based on domain generalization (DG) has gained growing attention due to its robustness in real-world applications. However, these DG methods rely heavily on labeled source data, which are usually costly and hard to access. Comparably, unlabeled face data are far more accessible in various scenarios. In this paper, we propose the first Unsupervised Domain Generalization framework for Face Anti-Spoofing, namely UDG-FAS, which could exploit large amounts of easily accessible unlabeled data to learn generalizable features for enhancing the low-data regime of FAS. Yet without supervision signals, le

arning intrinsic live/spoof features from complicated facial information is chal lenging, which is even tougher in cross-domain scenarios due to domain shift. Ex isting unsupervised learning methods tend to learn identity-biased and domain-bi ased features as shortcuts, and fail to specify spoof cues. To this end, we prop ose a novel Split-Rotation-Merge module to build identity-agnostic local represe ntations for mining intrinsic spoof cues and search the nearest neighbors in the same domain as positives for mitigating the identity bias. Moreover, we propose to search cross-domain neighbors with domain-specific normalization and merged local features to learn a domain-invariant feature space. To our best knowledge, this is the first attempt to learn generalized FAS features in a fully unsuperv ised way. Extensive experiments show that UDG-FAS significantly outperforms stat e-of-the-art methods on six diverse practical protocols.
****
DR-Tune: Improving Fine-tuning of Pretrained Visual Models by Distribution Regul arization with Semantic Calibration
Nan Zhou, Jiaxin Chen, Di Huang; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2023, pp. 1547-1556
The visual models pretrained on large-scale benchmarks encode general knowledge and prove effective in building more powerful representations for downstream tas ks. Most existing approaches follow the fine-tuning paradigm, either by initiali zing or regularizing the downstream model based on the pretrained one. The forme r fails to retain the knowledge in the successive fine-tuning phase, thereby pro ne to be over-fitting, and the latter imposes strong constraints to the weights or feature maps of the downstream model without considering semantic drift, ofte n incurring insufficient optimization. To deal with these issues, we propose a n ovel fine-tuning framework, namely distribution regularization with semantic cal ibration (DR-Tune). It employs distribution regularization by enforcing the down stream task head to decrease its classification error on the pretrained feature distribution, which prevents it from over-fitting while enabling sufficient trai ning of downstream encoders. Furthermore, to alleviate the interference by seman tic drift, we develop the semantic calibration (SC) module to align the global s hape and class centers of the pretrained and downstream feature distributions. E xtensive experiments on widely used image classification datasets show that DR-T une consistently improves the performance when combing with various backbones un der different pretraining strategies. Code is available at: https://github.com/w eeknan/DR-Tune.
****
MotionDeltaCNN: Sparse CNN Inference of Frame Differences in Moving Camera Video s with Spherical Buffers and Padded Convolutions
Mathias Parger, Chengcheng Tang, Thomas Neff, Christopher D. Twigg, Cem Keskin, Robert Wang, Markus Steinberger; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2023, pp. 17292-17301
Convolutional neural network inference on video input is computationally expensi ve and requires high memory bandwidth. Recently, DeltaCNN managed to reduce the cost by only processing pixels with significant updates over the previous frame. However, DeltaCNN relies on static camera input. Moving cameras add new challen ges in how to fuse newly unveiled image regions with already processed regions e fficiently to minimize the update rate - without increasing memory overhead and without knowing the camera extrinsics of future frames. In this work, we propose MotionDeltaCNN, a sparse CNN inference framework that supports moving cameras. We introduce spherical buffers and padded convolutions to enable seamless fusion of newly unveiled regions and previously processed regions - without increasing memory footprint. Our evaluation shows that we outperform DeltaCNN by up to 90% for moving camera videos.
****
General Image-to-Image Translation with One-Shot Image Guidance
Bin Cheng, Zuhao Liu, Yunbo Peng, Yue Lin; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2023, pp. 22736-22746
Large-scale text-to-image models pre-trained on massive text-image pairs show ex cellent performance in image synthesis recently. However, image can provide more

intuitive visual concepts than plain text. People may ask: how can we integrate the desired visual concept into an existing image, such as our portrait? Current methods are inadequate in meeting this demand as they lack the ability to preserve content or translate visual concepts effectively. Inspired by this, we propose a novel framework named visual concept translator (VCT) with the ability to preserve content in the source image and translate the visual concepts guided by a single reference image. The proposed VCT contains a content-concept inversion (CCI) process to extract contents and concepts, and a content-concept fusion (CCF) process to gather the extracted information to obtain the target image. Given only one reference image, the proposed VCT can complete a wide range of general image-to-image translation tasks with excellent results. Extensive experiments are conducted to prove the superiority and effectiveness of the proposed methods. Codes are available at https://github.com/CrystalNeuro/visual-concept-translator.
****
Dense 2D-3D Indoor Prediction with Sound via Aligned Cross-Modal Distillation
Heeseung Yun, Joonil Na, Gunhee Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7863-7872
Sound can convey significant information for spatial reasoning in our daily lives. To endow deep networks with such ability, we address the challenge of dense indoor prediction with sound in both 2D and 3D via cross-modal knowledge distillation. In this work, we propose a Spatial Alignment via Matching (SAM) distillation framework that elicits local correspondence between the two modalities in vision-to-audio knowledge transfer. SAM integrates audio features with visually coherent learnable spatial embeddings to resolve inconsistencies in multiple layers of a student model. Our approach does not rely on a specific input representation, allowing for flexibility in the input shapes or dimensions without performance degradation. With a newly curated benchmark named Dense Auditory Prediction of Surroundings (DAPS), we are the first to tackle dense indoor prediction of omnidirectional surroundings in both 2D and 3D with audio observations. Specifically, for audio-based depth estimation, semantic segmentation, and challenging 3D scene reconstruction, the proposed distillation framework consistently achieves state-of-the-art performance across various metrics and backbone architectures.
****
Leveraging SE(3) Equivariance for Learning 3D Geometric Shape Assembly
Ruihai Wu, Chenrui Tie, Yushi Du, Yan Zhao, Hao Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14311-14320
Shape assembly aims to reassemble parts (or fragments) into a complete object, which is a common task in our daily life. Different from the semantic part assembly (e.g., assembling a chair's semantic parts like legs into a whole chair), geometric part assembly (e.g., assembling bowl fragments into a complete bowl) is an emerging task in computer vision and robotics. Instead of semantic information, this task focuses on geometric information of parts. As the both geometric and pose space of fractured parts are exceptionally large, shape pose disentanglement of part representations is beneficial to geometric shape assembly. In our paper, we propose to leverage SE(3) equivariance for such shape pose disentanglement. Moreover, while previous works in vision and robotics only consider SE(3) equivariance for the representations of single objects, we move a step forward and propose leveraging SE(3) equivariance for representations considering multi-part correlations, which further boosts the performance of the multi-part assembly. Experiments demonstrate the significance of SE(3) equivariance and our proposed method for geometric shape assembly.
****
Adversarial Bayesian Augmentation for Single-Source Domain Generalization
Sheng Cheng, Tejas Gokhale, Yezhou Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11400-11410
Generalizing to unseen image domains is a challenging problem primarily due to the lack of diverse training data, inaccessible target data, and the large domain shift that may exist in many real-world settings. As such data augmentation is a critical component of domain generalization methods that seek to address this

problem. We present Adversarial Bayesian Augmentation (ABA), a novel algorithm that learns to generate image augmentations in the challenging single-source domain generalization setting. ABA draws on the strengths of adversarial learning and Bayesian neural networks to guide the generation of diverse data augmentations - these synthesized image domains aid the classifier in generalizing to unseen domains. We demonstrate the strength of ABA on several types of domain shift including style shift, subpopulation shift, and shift in the medical imaging setting. ABA outperforms all previous state-of-the-art methods, including pre-specified augmentations, pixel-based and convolutional-based augmentations. Code: https://github.com/shengcheng/ABA.
****
Robust Geometry-Preserving Depth Estimation Using Differentiable Rendering
Chi Zhang, Wei Yin, Gang Yu, Zhibin Wang, Tao Chen, Bin Fu, Joey Tianyi Zhou, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8951-8961
In this study, we address the challenge of 3D scene structure recovery from monocular depth estimation. While traditional depth estimation methods leverage labeled datasets to directly predict absolute depth, recent advancements advocate for mix-dataset training, enhancing generalization across diverse scenes. However, such mixed dataset training yields depth predictions only up to an unknown scale and shift, hindering accurate 3D reconstructions. Existing solutions necessitate extra 3D datasets or geometry-complete depth annotations, constraints that limit their versatility. In this paper, we propose a learning framework that trains models to predict geometry-preserving depth without requiring extra data or annotations. To produce realistic 3D structures, we render novel views of the reconstructed scenes and design loss functions to promote depth estimation consistency across different views. Comprehensive experiments underscore our framework's superior generalization capabilities, surpassing existing state-of-the-art methods on several benchmark datasets without leveraging extra training information. Moreover, our innovative loss functions empower the model to autonomously recover domain-specific scale-and-shift coefficients using solely unlabeled images.
****
Self-regulating Prompts: Foundational Model Adaptation without Forgetting
Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15190-15200
Prompt learning has emerged as an efficient alternative for fine-tuning foundational models, such as CLIP, for various downstream tasks. Conventionally trained using the task-specific objective, i.e., cross-entropy loss, prompts tend to overfit downstream data distributions and find it challenging to capture task-agnostic general features from the frozen CLIP. This leads to the loss of the model's original generalization capability. To address this issue, our work introduces a self-regularization framework for prompting called PromptSRC (Prompting with Self-regulating Constraints). PromptSRC guides the prompts to optimize for both task-specific and task-agnostic general representations using a three-pronged approach by: (a) regulating prompted representations via mutual agreement maximization with the frozen model, (b) regulating with self-ensemble of prompts over the training trajectory to encode their complementary strengths, and (c) regulating with textual diversity to mitigate sample diversity imbalance with the visual branch. To the best of our knowledge, this is the first regularization framework for prompt learning that avoids overfitting by jointly attending to pre-trained model features, the training trajectory during prompting, and the textual diversity. PromptSRC explicitly steers the prompts to learn a representation space that maximizes performance on downstream tasks without compromising CLIP generalization. We perform extensive experiments on 4 benchmarks where PromptSRC overall performs favorably well compared to the existing methods. Our code and pre-trained models are publicly available.
****
ASM: Adaptive Skinning Model for High-Quality 3D Face Modeling
Kai Yang, Hong Shang, Tianyang Shi, Xinghan Chen, Jingkai Zhou, Zhongqian Sun, W

ei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20708-20717

The research fields of parametric face model and 3D face reconstruction have been extensively studied. However, a critical question remains unanswered: how to tailor the face model for specific reconstruction settings. We argue that reconstruction with multi-view uncalibrated images demands a new model with stronger capacity. Our study shifts attention from data-dependent 3D Morphable Models (3DMM) to an understudied human-designed skinning model. We propose Adaptive Skinning Model (ASM), which redefines the skinning model with more compact and fully tunable parameters. With extensive experiments, we demonstrate that ASM achieves significantly improved capacity than 3DMM, with the additional advantage of model size and easy implementation for new topology. We achieve state-of-the-art performance with ASM for multi-view reconstruction on the Florence MICC Coop benchmark. Our quantitative analysis demonstrates the importance of a high-capacity model for fully exploiting abundant information from multi-view input in reconstruction. Furthermore, our model with physical-semantic parameters can be directly utilized for real-world applications, such as in-game avatar creation. As a result, our work opens up new research direction for parametric face model and facilitates future research on multi-view reconstruction.

****

EverLight: Indoor-Outdoor Editable HDR Lighting Estimation

Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, Jean-François Lalonde; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7420-7429

Because of the diversity in lighting environments, existing illumination estimation techniques have been designed explicitly on indoor or outdoor environments. Methods have focused specifically on capturing accurate energy (e.g., through parametric lighting models), which emphasizes shading and strong cast shadows; or producing plausible texture (e.g., with GANs), which prioritizes plausible reflections. Approaches which provide editable lighting capabilities have been proposed, but these tend to be with simplified lighting models, offering limited realism. In this work, we propose to bridge the gap between these recent trends in the literature, and propose a method which combines a parametric light model with 360deg panoramas, ready to use as HDRI in rendering engines. We leverage recent advances in GAN-based LDR panorama extrapolation from a regular image, which we extend to HDR using parametric spherical gaussians. To achieve this, we introduce a novel lighting co-modulation method that injects lighting-related features throughout the generator, tightly coupling the original or edited scene illumination within the panorama generation process. In our representation, users can easily edit light direction, intensity, number, etc. to impact shading while providing rich, complex reflections while seamlessly blending with the edits. Furthermore, our method encompasses indoor and outdoor environments, demonstrating state-of-the-art results even when compared to domain-specific methods.

****

MARS: Model-agnostic Biased Object Removal without Additional Supervision for Weakly-Supervised Semantic Segmentation

Sanghyun Jo, In-Jae Yu, Kyungsu Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 614-623

Weakly-supervised semantic segmentation aims to reduce labeling costs by training semantic segmentation models using weak supervision, such as image-level class labels. However, most approaches struggle to produce accurate localization maps and suffer from false predictions in class-related backgrounds (i.e., biased objects), such as detecting a railroad with the train class. Recent methods that remove biased objects require additional supervision for manually identifying biased objects for each problematic class and collecting their datasets by reviewing predictions, limiting their applicability to the real-world dataset with multiple labels and complex relationships for biasing. Following the first observation that biased features can be separated and eliminated by matching biased objects with backgrounds in the same dataset, we propose a fully-automatic/model-agnostic biased removal framework called MARS (Model-Agnostic biased object Removal w

ithout additional Supervision), which utilizes semantically consistent features
of an unsupervised technique to eliminate biased objects in pseudo labels. Surpr
isingly, we show that MARS achieves new state-of-the-art results on two popular
benchmarks, PASCAL VOC 2012 (val: 77.7%, test: 77.2%) and MS COCO 2014 (val: 49.
4%), by consistently improving the performance of various WSSS models by at leas
t 30% without additional supervision. Code is available at https://github.com/sh
jo-april/MARS.
****
CAFA: Class-Aware Feature Alignment for Test-Time Adaptation
Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, Jaegul Choo
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2023, pp. 19060-19071
Despite recent advancements in deep learning, deep neural networks continue to s
uffer from performance degradation when applied to new data that differs from tr
aining data. Test-time adaptation (TTA) aims to address this challenge by adapti
ng a model to unlabeled data at test time. TTA can be applied to pretrained netw
orks without modifying their training procedures, enabling them to utilize a wel
l-formed source distribution for adaptation. One possible approach is to align t
he representation space of test samples to the source distribution (i.e., featur
e alignment). However, performing feature alignment in TTA is especially challen
ging in that access to labeled source data is restricted during adaptation. That
is, a model does not have a chance to learn test data in a class-discriminative
manner, which was feasible in other adaptation tasks (e.g., unsupervised domain
adaptation) via supervised losses on the source data. Based on this observation
, we propose a simple yet effective feature alignment loss, termed as Class-Awar
e Feature Alignment (CAFA), which simultaneously 1) encourages a model to learn
target representations in a class-discriminative manner and 2) effectively mitig
ates the distribution shifts at test time. Our method does not require any hyper
-parameters or additional losses, which are required in previous approaches. We
conduct extensive experiments on 6 different datasets and show our proposed meth
od consistently outperforms existing baselines.
****
Learning Clothing and Pose Invariant 3D Shape Representation for Long-Term Perso
n Re-Identification
Feng Liu, Minchul Kim, ZiAng Gu, Anil Jain, Xiaoming Liu; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19617-19626
Long-Term Person Re-Identification (LT-ReID) has become increasingly crucial in
computer vision and biometrics. In this work, we aim to extend LT-ReID beyond pe
destrian recognition to include a wider range of real-world human activities whi
le still accounting for cloth-changing scenarios over large time gaps. This sett
ing poses additional challenges due to the geometric misalignment and appearance
ambiguity caused by the diversity of human pose and clothing. To address these
challenges, we propose a new approach 3DInvarReID for (i) disentangling identity
from non-identity components (pose, clothing shape, and texture) of 3D clothed
humans, and (ii) reconstructing accurate 3D clothed body shapes and learning dis
criminative features of naked body shapes for person ReID in a joint manner. To
better evaluate our study of LT-ReID, we collect a real-world dataset called CCD
A, which contains a wide variety of human activities and clothing changes. Exper
imentally, we show the superior performance of our approach for person ReID.
****
Agile Modeling: From Concept to Classifier in Minutes
Otilia Stretcu, Edward Vendrow, Kenji Hata, Krishnamurthy Viswanathan, Vittorio
Ferrari, Sasan Tavakkol, Wenlei Zhou, Aditya Avinash, Emming Luo, Neil Gordon Al
ldrin, MohammadHossein Bateni, Gabriel Berger, Andrew Bunner, Chun-Ta Lu, Javier
Rey, Giulia DeSalvo, Ranjay Krishna, Ariel Fuxman■; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 22323-22334
The application of computer vision methods to nuanced, subjective concepts is gr
owing. While crowdsourcing has served the vision community well for most objecti
ve tasks (such as labeling a "zebra"), it now falters on tasks where there is su
bstantial subjectivity in the concept (such as identifying "gourmet tuna"). Howe

ver, empowering any user to develop a classifier for their concept is technicall y difficult: users are neither machine learning experts nor have the patience to label thousands of examples. In reaction, we introduce the problem of Agile Mod eling: the process of turning any subjective visual concept into a computer visi on model through a real-time user-in-the-loop interactions. We instantiate an Ag ile Modeling prototype for image classification and show through a user study (N =14) that users can create classifiers with minimal effort under 30 minutes. We compare this user driven process with the traditional crowdsourcing paradigm and find that the crowd's notion often differs from that of the user's, especially as the concepts become more subjective. Finally, we scale our experiments with s imulations of users training classifiers for ImageNet21k categories to further d emonstrate the efficacy.
****
Improving Lens Flare Removal with General-Purpose Pipeline and Multiple Light So urces Recovery
Yuyan Zhou, Dong Liang, Songcan Chen, Sheng-Jun Huang, Shuo Yang, Chongyi Li; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 23, pp. 12969-12979
When taking images against strong light sources, the resulting images often cont ain heterogeneous flare artifacts. These artifacts can importantly affect image visual quality and downstream computer vision tasks. While collecting real data pairs of flare-corrupted/flare-free images for training flare removal models is challenging, current methods utilize the direct-add approach to synthesize data.

However, these methods do not consider automatic exposure and tone mapping in i mage signal processing pipeline (ISP), leading to the limited generalization cap ability of deep models training using such data. Besides, existing methods strug gle to handle multiple light sources due to the different sizes, shapes and illu minance of various light sources. In this paper, we propose a solution to improv e the performance of lens flare removal by revisiting the ISP and remodeling the principle of automatic exposure in the synthesis pipeline and design a more rel iable light sources recovery strategy. The new pipeline approaches realistic ima ging by discriminating the local and global illumination through convex combinat ion, avoiding global illumination shifting and local over-saturation. Our strate gy for recovering multiple light sources convexly averages the input and output of the neural network based on illuminance levels, thereby avoiding the need for a hard threshold in identifying light sources. We also contribute a new flare r emoval testing dataset containing the flare-corrupted images captured by ten typ es of consumer electronics. The dataset facilitates the verification of the gene ralization capability of flare removal methods. Extensive experiments show that our solution can effectively improve the performance of lens flare removal and p ush the frontier toward more general situations.
****
FACET: Fairness in Computer Vision Evaluation Benchmark
Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng -Yang Fu, Melissa Hall, Candace Ross; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20370-20382
Computer vision models have known performance disparities across attributes such as gender and skin tone. This means during tasks such as classification and det ection, model performance differs for certain classes based on the demographics of the people in the image. These disparities have been shown to exist, but unti l now there has not been a unified approach to measure these differences for com mon use-cases of computer vision models. We present a new benchmark named FACET (FAirness in Computer Vision EvaluaTion), a large, publicly available evaluation set of 32k images for some of the most common vision tasks - image classificati on, object detection and segmentation. For every image in FACET, we hired expert reviewers to manually annotate person-related attributes such as perceived skin tone and hair type, manually draw bounding boxes and label fine-grained person- related classes such as disk jockey or guitarist. In addition, we use FACET to b enchmark state-of-the-art vision models and present a deeper understanding of po

tential performance disparities and challenges across sensitive demographic attr
ibutes. With the exhaustive annotations collected, we probe models using single
demographics attributes as well as multiple attributes using an intersectional a
pproach (e.g. hair color and perceived skin tone). Our results show that classif
ication, detection, segmentation, and visual grounding models exhibit performanc
e disparities across demographic attributes and intersections of attributes. The
se harms suggest that not all people represented in datasets receive fair and eq
uitable treatment in these vision tasks. We hope current and future results usin
g our benchmark will contribute
to fairer, more robust vision models. FACET is available publicly at https://fa
cet.metademolab.com
****
Few-Shot Physically-Aware Articulated Mesh Generation via Hierarchical Deformati
on
Xueyi Liu, Bin Wang, He Wang, Li Yi; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 854-864
We study the problem of few-shot physically-aware articulated mesh generation. B
y observing an articulated object dataset containing only a few examples, we wis
h to learn a model that can generate diverse meshes with high visual fidelity an
d physical validity. Previous mesh generative models either have difficulties in
depicting a diverse data space from only a few examples or fail to ensure physi
cal validity of their samples. Regarding the above challenges, we propose two ke
y innovations, including 1) a hierarchical mesh deformation-based generative mod
el based upon the divide-and-conquer philosophy to alleviate the few-shot challe
nge by borrowing transferrable deformation patterns from large scale rigid meshe
s and 2) a physics-aware deformation correction scheme to encourage physically p
lausible generations. We conduct extensive experiments on 6 articulated categori
es to demonstrate the superiority of our method in generating articulated meshes
with better diversity, higher visual fidelity, and better physical validity ove
r previous methods in the few-shot setting. Further, we validate solid contribut
ions of our two innovations in the ablation study. Project page with code is ava
ilable at https://meowuu7.github.io/few-arti-obj-gen.
****
Single-Stage Diffusion NeRF: A Unified Approach to 3D Generation and Reconstruct
ion
Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, Hao Su;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2023, pp. 2416-2425
3D-aware image synthesis encompasses a variety of tasks, such as scene generatio
n and novel view synthesis from images. Despite numerous task-specific methods,
developing a comprehensive model remains challenging. In this paper, we present
SSDNeRF, a unified approach that employs an expressive diffusion model to learn
a generalizable prior of neural radiance fields (NeRF) from multi-view images of
diverse objects. Previous studies have used two-stage approaches that rely on p
retrained NeRFs as real data to train diffusion models. In contrast, we propose
a new single-stage training paradigm with an end-to-end objective that jointly o
ptimizes a NeRF auto-decoder and a latent diffusion model, enabling simultaneous
3D reconstruction and prior learning, even from sparsely available views. At te
st time, we can directly sample the diffusion prior for unconditional generation
, or combine it with arbitrary observations of unseen objects for NeRF reconstru
ction. SSDNeRF demonstrates robust results comparable to or better than leading
task-specific methods in unconditional generation and single/sparse-view 3D reco
nstruction.
****
DCPB: Deformable Convolution Based on the Poincare Ball for Top-view Fisheye Cam
eras
Xuan Wei, Zhidan Ran, Xiaobo Lu; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2023, pp. 13308-13317
The accuracy of the visual tasks for top-view fisheye cameras is limited by the
Euclidean geometry for pose-distorted objects in images. In this paper, we demon

strate the analogy between the fisheye model and the Poincare ball and that lear ning the shape of convolution kernels in the Poincare Ball can alleviate the spa tial distortion problem. In particular, we propose the Deformable Convolution ba sed on the Poincare Ball, named DCPB, which conducts the Graph Convolutional Net work (GCN) in the Poincare ball and calculates the geodesic distances to Poincar e hyperplanes as the offsets and modulation scalars of the modulated deformable convolution. Besides, we explore an appropriate network structure in the baselin e with the DCPB. The DCPB markedly improves the neural network's performance. Ex perimental results on the public dataset THEODORE show that DCPB obtains a highe r accuracy, and its efficiency demonstrates the potential for using temporal inf ormation in fisheye videos.
****
Integrating Boxes and Masks: A Multi-Object Framework for Unified Visual Trackin g and Segmentation
Yuanyou Xu, Zongxin Yang, Yi Yang; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2023, pp. 9738-9751
Tracking any given object(s) spatially and temporally is a common purpose in Vis ual Object Tracking (VOT) and Video Object Segmentation (VOS). Joint tracking an d segmentation have been attempted in some studies but they often lack full comp atibility of both box and mask in initialization and prediction, and mainly focu s on single-object scenarios. To address these limitations, this paper proposes a Multi-object Mask-box Integrated framework for unified Tracking and Segmentati on, dubbed MITS. Firstly, the unified identification module is proposed to suppo rt both box and mask reference for initialization, where detailed object informa tion is inferred from boxes or directly retained from masks. Additionally, a nov el pinpoint box predictor is proposed for accurate multi-object box prediction, facilitating target-oriented representation learning. All target objects are pro cessed simultaneously from encoding to propagation and decoding, as a unified pi peline for VOT and VOS. Experimental results show MITS achieves state-of-the-art performance on both VOT and VOS benchmarks. Notably, MITS surpasses the best pr ior VOT competitor by around 6% on the GOT-10k test set, and significantly impro ves the performance of box initialization on VOS benchmarks. The code is availab le at https://github.com/yoxu515/MITS.
****
One-Shot Generative Domain Adaptation
Ceyuan Yang, Yujun Shen, Zhiyi Zhang, Yinghao Xu, Jiapeng Zhu, Zhirong Wu, Bolei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7733-7742
This work aims to transfer a Generative Adversarial Network (GAN) pre-trained on one image domain to another domain referred to as few as just one reference ima ge. The challenge is that, under limited supervision, it is extremely difficult to synthesize photo realistic and highly diverse images while retaining the repr esentative characters of the target domain. Different from existing approaches t hat adopt the vanilla fine-tuning strategy, we design two lightweight modules in the generator and the discriminator respectively. We first introduce an attribu te adaptor in the generator and freeze the generator's original parameters, whic h can reuse the prior knowledge to the most extent and maintain the synthesis qu ality and diversity. We then equip the well-learned discriminator with an attrib ute classifier to ensure that the generator with the attribute adaptor captures the appropriate characters of the reference image. Furthermore, considering the very limited diversity of the training data (i.e., as few as only one image), we propose to constrain the diversity of the latent space through truncation in th e training process, alleviating the optimization difficulty. Our approach brings appealing results under various settings, substantially surpassing state-of-the -art alternatives, especially in terms of synthesis diversity. Noticeably, our m ethod works well even with large domain gaps and robustly converges within a few minutes for each experiment. Code and models are available at https://genforce. github.io/genda/.
****
Prototypes-oriented Transductive Few-shot Learning with Conditional Transport

Long Tian, Jingyi Feng, Xiaoqiang Chai, Wenchao Chen, Liming Wang, Xiyang Liu, Bo Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16317-16326

Transductive Few-Shot Learning (TFSL) has recently attracted increasing attention since it typically outperforms its inductive peer by leveraging statistics of query samples.However, previous TFSL methods usually encode uniform prior that all the classes within query samples are equally likely, which is biased in imbalanced TFSL and causes severe performance degradation.Given this pivotal issue, in this work, we propose a novel Conditional Transport (CT) based imbalanced TFSL model called Prototypes-oriented Unbiased Transfer Model (PUTM) to fully exploit unbiased statistics of imbalanced query samples, which employs forward and backward navigators as transport matrices to balance the prior of query samples per class between uniform and adaptive data-driven distributions. For efficiently transferring statistics learned by CT, we further derive a closed form solution to refine prototypes based on MAP given the learned navigators. The above two steps of discovering and transferring unbiased statistics follow an iterative manner, formulating our EM-based solver. Experimental results on four standard benchmarks including miniImageNet, tieredImageNet, CUB, and CIFAR-FS demonstrate superiority of our model in class-imbalanced generalization.

****

SparseFusion: Fusing Multi-Modal Sparse Representations for Multi-Sensor 3D Object Detection

Yichen Xie, Chenfeng Xu, Marie-Julie Rakotosaona, Patrick Rim, Federico Tombari, Kurt Keutzer, Masayoshi Tomizuka, Wei Zhan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17591-17602

By identifying four important components of existing LiDAR-camera 3D object detection methods (LiDAR and camera candidates, transformation, and fusion outputs), we observe that all existing methods either find dense candidates or yield dense representations of scenes. However, given that objects occupy only a small part of a scene, finding dense candidates and generating dense representations is noisy and inefficient. We propose SparseFusion, a novel multi-sensor 3D detection method that exclusively uses sparse candidates and sparse representations. Specifically, SparseFusion utilizes the outputs of parallel detectors in the LiDAR and camera modalities as sparse candidates for fusion. We transform the camera candidates into the LiDAR coordinate space by disentangling the object representations. Then, we can fuse the multi-modality candidates in a unified 3D space by a lightweight self-attention module. To mitigate negative transfer between modalities, we propose novel semantic and geometric cross-modality transfer modules that are applied prior to the modality-specific detectors. SparseFusion achieves state-of-the-art performance on the nuScenes benchmark while also running at the fastest speed, even outperforming methods with stronger backbones. We perform extensive experiments to demonstrate the effectiveness and efficiency of our modules and overall method pipeline. Our code will be made publicly available at https://github.com/yichen928/SparseFusion.

****

DetermiNet: A Large-Scale Diagnostic Dataset for Complex Visually-Grounded Referencing using Determiners

Clarence Lee, M Ganesh Kumar, Cheston Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20019-20028

State-of-the-art visual grounding models can achieve high detection accuracy, but they are not designed to distinguish between all objects versus only certain objects of interest. In natural language, in order to specify a particular object or set of objects of interest, humans use determiners such as "my", "either" and "those". Determiners, as an important word class, are a type of schema in natural language about the reference or quantity of the noun. Existing grounded referencing datasets place much less emphasis on determiners, compared to other word classes such as nouns, verbs and adjectives. This makes it difficult to develop models that understand the full variety and complexity of object referencing. Thus, we have developed and released the DetermiNet dataset, which comprises 250,000 synthetically generated images and captions based on 25 determiners. The tas

k is to predict bounding boxes to identify objects of interest, constrained by the semantics of the given determiner. We find that current state-of-the-art visual grounding models do not perform well on the dataset, highlighting the limitations of existing models on reference and quantification tasks.
****
3DMOTFormer: Graph Transformer for Online 3D Multi-Object Tracking
Shuxiao Ding, Eike Rehder, Lukas Schneider, Marius Cordts, Juergen Gall; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9784-9794
Tracking 3D objects accurately and consistently is crucial for autonomous vehicles, enabling more reliable downstream tasks such as trajectory prediction and motion planning. Based on the substantial progress in object detection in recent years, the tracking-by-detection paradigm has become a popular choice due to its simplicity and efficiency. State-of-the-art 3D multi-object tracking (MOT) approaches typically rely on non-learned model-based algorithms such as Kalman Filter but require many manually tuned parameters. On the other hand, learning-based approaches face the problem of adapting the training to the online setting, leading to inevitable distribution mismatch between training and inference as well as suboptimal performance. In this work, we propose 3DMOTFormer, a learned geometry-based 3D MOT framework building upon the transformer architecture. We use an Edge-Augmented Graph Transformer to reason on the track-detection bipartite graph frame-by-frame and conduct data association via edge classification. To reduce the distribution mismatch between training and inference, we propose a novel online training strategy with an autoregressive and recurrent forward pass as well as sequential batch optimization. Using CenterPoint detections, our approach achieves 71.2% and 68.2% AMOTA on the nuScenes validation and test split, respectively. In addition, a trained 3DMOTFormer model generalizes well across different object detectors. Code is available at: https://github.com/dsx0511/3DMOTFormer.
****
ReGen: A good Generative Zero-Shot Video Classifier Should be Rewarded
Adrian Bulat, Enrique Sanchez, Brais Martinez, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13523-13533
This paper sets out to solve the following problem: How can we turn a generative video captioning model into an open-world video/action classification model? Video captioning models can naturally produce open-ended free-form descriptions of a given video which, however, might not be discriminative enough for video/action recognition. Unfortunately, when fine-tuned to auto-regress the class names directly, video captioning models overfit the base classes losing their open-world zero-shot capabilities. To alleviate base class overfitting, in this work, we propose to use reinforcement learning to enforce the output of the video captioning model to be more class-level discriminative. Specifically, we propose ReGen, a novel reinforcement learning based framework with a three-fold objective and reward functions: (1) a class-level discrimination reward that enforces the generated caption to be correctly classified into the corresponding action class, (2) a CLIP reward that encourages the generated caption to continue to be descriptive of the input video (i.e. video-specific), and (3) a grammar reward that preserves the grammatical correctness of the caption. We show that ReGen can train a model to produce captions that are: discriminative, video-specific and grammatically correct. Importantly, when evaluated on standard benchmarks for zero- and few-shot action classification, ReGen significantly outperforms the previous state-of-the-art.
****
Complementary Domain Adaptation and Generalization for Unsupervised Continual Domain Shift Learning
Wonguk Cho, Jinha Park, Taesup Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11442-11452
Continual domain shift poses a significant challenge in real-world applications, particularly in situations where labeled data is not available for new domains. The challenge of acquiring knowledge in this problem setting is referred to as

unsupervised continual domain shift learning. Existing methods for domain adapta
tion and generalization have limitations in addressing this issue, as they focus
 either on adapting to a specific domain or generalizing to unseen domains, but
not both. In this paper, we propose Complementary Domain Adaptation and Generali
zation (CoDAG), a simple yet effective learning framework that combines domain a
daptation and generalization in a complementary manner to achieve three major go
als of unsupervised continual domain shift learning: adapting to a current domai
n, generalizing to unseen domains, and preventing forgetting of previously seen
domains. Our approach is model-agnostic, meaning that it is compatible with any
existing domain adaptation and generalization algorithms. We evaluate CoDAG on s
everal benchmark datasets and demonstrate that our model outperforms state-of-th
e-art models in all datasets and evaluation metrics, highlighting its effectiven
ess and robustness in handling unsupervised continual domain shift learning.
****
RICO: Regularizing the Unobservable for Indoor Compositional Reconstruction
Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, Yong Liu; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
3, pp. 17761-17771
Recently, neural implicit surfaces have become popular for multi-view reconstruc
tion. To facilitate practical applications like scene editing and manipulation,
some works extend the framework with semantic masks input for the object-composi
tional reconstruction rather than the holistic perspective. Though achieving pla
usible disentanglement, the performance drops significantly when processing the
indoor scenes where objects are usually partially observed. We propose RICO to a
ddress this by regularizing the unobservable regions for indoor compositional re
construction. Our key idea is to first regularize the smoothness of the occluded
 background, which then in turn guides the foreground object reconstruction in u
nobservable regions based on the object-background relationship. Particularly, w
e regularize the geometry smoothness of occluded background patches. With the im
proved background surface, the signed distance function and the reversedly rende
red depth of objects can be optimized to bound them within the background range.
 Extensive experiments show our method outperforms other methods on synthetic an
d real-world indoor scenes and prove the effectiveness of proposed regularizatio
ns. The code is available at https://github.com/kyleleey/RICO.
****
Ordered Atomic Activity for Fine-grained Interactive Traffic Scenario Understand
ing
Nakul Agarwal, Yi-Ting Chen; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2023, pp. 8624-8636
We introduce a novel representation called Ordered Atomic Activity for interacti
ve scenario understanding. The representation decomposes each scenario into a se
t of ordered atomic activities, where each activity consists of an action and th
e corresponding actors involved and the order denotes the temporal development o
f the scenario. The design also helps in identifying important interactive relat
ionships such as yielding. The action is a high-level semantic motion pattern th
at is grounded in the surrounding road topology, which we decompose into zones a
nd corners with unique IDs. For example, a group of pedestrians crossing on the
left side is denoted as C1 - C4: P+, as depicted in Figure 1. We collect a new l
arge-scale dataset called OATS (Ordered Atomic Activities in interactive Traffic
 Scenarios), comprising 1026 video clips (20s) captured at intersections. Each
 clip is labeled with the proposed language, resulting in 59 activity categories
 and 6512 annotated activity instances. We propose three fine-grained scenario u
nderstanding tasks, i.e., multi-label Atomic Activity recognition, recognition,
activity order prediction, and interactive scenario retrieval. We implement vari
ous state-of-the-art algorithms and conduct extensive experiments on OATS. We fo
und the existing methods cannot achieve satisfactory performance, indicating new
 opportunities for the community to develop new algorithms for these tasks towar
d better interactive scenario understanding
****
CO-PILOT: Dynamic Top-Down Point Cloud with Conditional Neighborhood Aggregation

for Multi-Gigapixel Histopathology Image Representation
Ramin Nakhli, Allen Zhang, Ali Mirabadi, Katherine Rich, Maryam Asadi, Blake Gil
ks, Hossein Farahani, Ali Bashashati; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2023, pp. 21063-21073
Predicting survival rates based on multi-gigapixel histopathology images is one
of the most challenging tasks in digital pathology. Due to the computational com
plexities, Multiple Instance Learning (MIL) has become the conventional approach
for this process as it breaks the image into smaller patches. However, this tec
hnique fails to account for the individual cells present in each patch, while th
ey are the fundamental part of the tissue. In this work, we developed a novel dy
namic and hierarchical point-cloud-based method (CO-PILOT) for the processing of
cellular graphs extracted from routine histopathology images. By using bottom-u
p information propagation and top-down conditional attention, our model gains ac
cess to an adaptive focus across different levels of tissue hierarchy. Through c
omprehensive experiments, we demonstrate that our model can outperform all the s
tate-of-the-art methods in survival prediction, including the hierarchical Visio
n Transformer (ViT), across two datasets and four metrics with only half of the
parameters of the closest baseline. Importantly, our model is able to stratify t
he patients into different risk cohorts with statistically different outcomes ac
ross two large datasets, a task that was previously achievable only using genomi
c information. Furthermore, we publish a large dataset containing 873 cellular g
raphs from 188 patients, along with their survival information, making it one of
the largest publicly available datasets in this context.
****
Troubleshooting Ethnic Quality Bias with Curriculum Domain Adaptation for Face I
mage Quality Assessment
Fu-Zhao Ou, Baoliang Chen, Chongyi Li, Shiqi Wang, Sam Kwong; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20718-20
729
Face Image Quality Assessment (FIQA) lays the foundation for ensuring the stabil
ity and accuracy of face recognition systems. However, existing FIQA methods mai
nly formulate quality relationships within the training set to yield quality sco
res, ignoring the generalization problem caused by ethnic quality bias between t
he training and test sets. Domain adaptation presents a potential solution to mi
tigate the bias, but if FIQA is treated essentially as a regression task, it wil
l be limited by the challenge of feature scaling in transfer learning. Additiona
lly, how to guarantee source risk is also an issue due to the lack of ground-tru
th labels of the source domain for FIQA. This paper presents the first attempt i
n the field of FIQA to address these challenges with a novel Ethnic-Quality-Bias
Mitigating (EQBM) framework. Specifically, to eliminate the restriction of scal
ar regression, we first compute the Likert-scale quality probability distributio
ns as source domain annotations. Furthermore, we design an easy-to-hard training
scheduler based on the inter-domain uncertainty and intra-domain quality margin
as well as the ranking-based domain adversarial network to enhance the effectiv
eness of transfer learning and further reduce the source risk in domain adaptati
on. Extensive experiments demonstrate that the EQBM significantly mitigates the
quality bias and improves the generalization capability of FIQA across races on
different datasets.
****
HybridAugment++: Unified Frequency Spectra Perturbations for Model Robustness
Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, Pinar Duygulu; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5718-5728
Convolutional Neural Networks (CNN) are known to exhibit poor generalization per
formance under distribution shifts. Their generalization have been studied exten
sively, and one line of work approaches the problem from a frequency-centric per
spective. These studies highlight the fact that humans and CNNs might focus on d
ifferent frequency components of an image. First, inspired by these observations
, we propose a simple yet effective data augmentation method HybridAugment that
reduces the reliance of CNNs on high-frequency components, and thus improves the
ir robustness while keeping their clean accuracy high. Second, we propose Hybrid

Augment++, which is a hierarchical augmentation method that attempts to unify various frequency-spectrum augmentations. HybridAugment++ builds on HybridAugment, and also reduces the reliance of CNNs on the amplitude component of images, and promotes phase information instead. This unification results in competitive to or better than state-of-the-art results on clean accuracy (CIFAR-10/100 and ImageNet), corruption benchmarks (ImageNet-C, CIFAR-10-C and CIFAR-100-C), adversarial robustness on CIFAR-10 and out-of-distribution detection on various datasets. HybridAugment
and HybridAugment++ are implemented in a few lines of code, does not require extra data, ensemble models or additional networks.
****
CLR: Channel-wise Lightweight Reprogramming for Continual Learning
Yunhao Ge, Yuecheng Li, Shuo Ni, Jiaping Zhao, Ming-Hsuan Yang, Laurent Itti; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18798-18808
Continual learning aims to emulate the human ability to continually accumulate knowledge over sequential tasks. The main challenge is to maintain performance on previously learned tasks after learning new tasks, i.e., to avoid catastrophic forgetting. We propose a Channel-wise Lightweight Reprogramming (CLR) approach that helps convolutional neural networks (CNNs) overcome catastrophic forgetting during continual learning. We show that a CNN model trained on an old task (or self-supervised proxy task) could be "reprogrammed" to solve a new task by using our proposed lightweight (very cheap) reprogramming parameter. With the help of CLR, we have a better stability-plasticity trade-off to solve continual learning problems: To maintain stability and retain previous task ability, we use a common task-agnostic immutable part as the shared "anchor" parameter set. We then add task-specific lightweight reprogramming parameters to reinterpret the outputs of the immutable parts, to enable plasticity and integrate new knowledge. To learn sequential tasks, we only train the lightweight reprogramming parameters to learn each new task. Reprogramming parameters are task-specific and exclusive to each task, which makes our method immune to catastrophic forgetting. To minimize the parameter requirement of reprogramming to learn new tasks, we make reprogramming lightweight by only adjusting essential kernels and learning channel-wise linear mappings from anchor parameters to task-specific domain knowledge. We show that, for general CNNs, the CLR parameter increase is less than 0.6% for any new task. Our method outperforms 13 state-of-the-art continual learning baselines on a new challenging sequence of 53 image classification datasets. Code and data are here: https://github.com/gyhandy/Channel-wise-Lightweight-Reprogramming
****
IOMatch: Simplifying Open-Set Semi-Supervised Learning with Joint Inliers and Outliers Utilization
Zekun Li, Lei Qi, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15870-15879
Semi-supervised learning (SSL) aims to leverage massive unlabeled data when labels are expensive to obtain. Unfortunately, in many real-world applications, the collected unlabeled data will inevitably contain unseen-class outliers not belonging to any of the labeled classes. To deal with the challenging open-set SSL task, the mainstream methods tend to first detect outliers and then filter them out. However, we observe a surprising fact that such approach could result in more severe performance degradation when labels are extremely scarce, as the unreliable outlier detector may wrongly exclude a considerable portion of valuable inliers. To tackle with this issue, we introduce a novel open-set SSL framework, IOMatch, which can jointly utilize inliers and outliers, even when it is difficult to distinguish exactly between them. Specifically, we propose to employ a multi-binary classifier in combination with the standard closed-set classifier for producing unified open-set classification targets, which regard all outliers as a single new class. By adopting these targets as open-set pseudo-labels, we optimize an open-set classifier with all unlabeled samples including both inliers and outliers. Extensive experiments have shown that IOMatch significantly outperforms the baseline methods across different benchmark datasets and different settings

despite its remarkable simplicity. Our code and models are available at https://github.com/nukezil/IOMatch.
****
Hierarchical Point-based Active Learning for Semi-supervised Point Cloud Semantic Segmentation
Zongyi Xu, Bo Yuan, Shanshan Zhao, Qianni Zhang, Xinbo Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18098-18108
Impressive performance on point cloud semantic segmentation has been achieved by fully-supervised methods with large amounts of labelled data. As it is labour-intensive to acquire large-scale point cloud data with point-wise labels, many attempts have been made to explore learning 3D point cloud segmentation with limited annotations. Active learning is one of the effective strategies to achieve this purpose but is still under-explored. The most recent methods of this kind measure the uncertainty of each pre-divided region for manual labelling but they suffer from redundant information and require additional efforts for region division. This paper aims at addressing this issue by developing a hierarchical point-based active learning strategy. Specifically, we measure the uncertainty for each point by a hierarchical minimum margin uncertainty module which considers the contextual information at multiple levels. Then, a feature-distance suppression strategy is designed to select important and representative points for manual labelling. Besides, to better exploit the unlabelled data, we build a semi-supervised segmentation framework based on our active strategy. Extensive experiments on the S3DIS and ScanNetV2 datasets demonstrate that the proposed framework achieves 96.5% and 100% performance of fully-supervised baseline with only 0.07% and 0.1% training data, respectively, outperforming the state-of-the-art weakly-supervised and active learning methods. The code will be available at https://github.com/SmiletoE/HPAL.
****
Doppelgangers: Learning to Disambiguate Images of Similar Structures
Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, Noah Snavely; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 34-44
We consider the visual disambiguation task of determining whether a pair of visually similar images depict the same or distinct 3D surfaces (e.g., the same or opposite sides of a symmetric building). Illusory image matches, where two images observe distinct but visually similar 3D surfaces, can be challenging for humans to differentiate, and can also lead 3D reconstruction algorithms to produce erroneous results. We propose a learning-based approach to visual disambiguation, formulating it as a binary classification task on image pairs. To that end, we introduce a new dataset for this problem, Doppelgangers, which includes image pairs of similar structures with ground truth labels. We also design a network architecture that takes the spatial distribution of local keypoints and matches as input, allowing for better reasoning about both local and global cues. Our evaluation shows that our method can distinguish illusory matches in difficult cases, and can be integrated into SfM pipelines to produce correct, disambiguated 3D reconstructions. See our project page for our code, datasets, and more results: http://doppelgangers-3d.github.io/.
****
BEV-DG: Cross-Modal Learning under Bird's-Eye View for Domain Generalization of 3D Semantic Segmentation
Miaoyu Li, Yachao Zhang, Xu Ma, Yanyun Qu, Yun Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11632-11642
Cross-modal Unsupervised Domain Adaptation (UDA) aims to exploit the complementarity of 2D-3D data to overcome the lack of annotation in a new domain. However, UDA methods rely on access to the target domain during training, meaning the trained model only works in a specific target domain. In light of this, we propose cross-modal learning under bird's-eye view for Domain Generalization (DG) of 3D semantic segmentation, called BEV-DG. DG is more challenging because the model cannot access the target domain during training, meaning it needs to rely on cros

s-modal learning to alleviate the domain gap. Since 3D semantic segmentation req
uires the classification of each point, existing cross-modal learning is directl
y conducted point-to-point, which is sensitive to the misalignment in projection
s between pixels and points. To this end, our approach aims to optimize domain-i
rrelevant representation modeling with the aid of cross-modal learning under bir
d's-eye view. We propose BEV-based Area-to-area Fusion (BAF) to conduct cross-mo
dal learning under bird's-eye view, which has a higher fault tolerance for point
-level misalignment. Furthermore, to model domain-irrelevant representations, we
 propose BEV-driven Domain Contrastive Learning (BDCL) with the help of cross-mo
dal learning under bird's-eye view. We design three domain generalization settin
gs based on three 3D datasets, and BEV-DG significantly outperforms state-of-the
-art competitors with tremendous margins in all settings.
****
Grounded Entity-Landmark Adaptive Pre-Training for Vision-and-Language Navigatio
n
Yibo Cui, Liang Xie, Yakun Zhang, Meishan Zhang, Ye Yan, Erwei Yin; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12
043-12053
Cross-modal alignment is one key challenge for Vision-and-Language Navigation (V
LN). Most existing studies concentrate on mapping the global instruction or sing
le sub-instruction to the corresponding trajectory. However, another critical pr
oblem of achieving fine-grained alignment at the entity level is seldom consider
ed. To address this problem, we propose a novel Grounded Entity-Landmark Adaptiv
e (GELA) pre-training paradigm for VLN tasks. To achieve the adaptive pre-traini
ng paradigm, we first introduce grounded entity-landmark human annotations into
the Room-to-Room (R2R) dataset, named GEL-R2R. Additionally, we adopt three grou
nded entity-landmark adaptive pre-training objectives: 1) entity phrase predicti
on, 2) landmark bounding box prediction, and 3) entity-landmark semantic alignme
nt, which explicitly supervise the learning of fine-grained cross-modal alignmen
t between entity phrases and environment landmarks. Finally, we validate our mod
el on two downstream benchmarks: VLN with descriptive instructions (R2R) and dia
logue instructions (CVDN). The comprehensive experiments show that our GELA mode
l achieves state-of-the-art results on both tasks, demonstrating its effectivene
ss and generalizability.
****
Lip Reading for Low-resource Languages by Learning and Combining General Speech
Knowledge and Language-specific Knowledge
Minsu Kim, Jeong Hun Yeo, Jeongsoo Choi, Yong Man Ro; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2023, pp. 15359-15371
This paper proposes a novel lip reading framework, especially for low-resource l
anguages, which has not been well addressed in the previous literature. Since lo
w-resource languages do not have enough video-text paired data to train the mode
l to have sufficient power to model lip movements and language, it is regarded a
s challenging to develop lip reading models for low-resource languages. In order
 to mitigate the challenge, we try to learn general speech knowledge, the abilit
y to model lip movements, from a high-resource language through the prediction o
f speech units. It is known that different languages partially share common phon
emes, thus general speech knowledge learned from one language can be extended to
 other languages. Then, we try to learn language-specific knowledge, the ability
 to model language, by proposing Language-specific Memory-augmented Decoder (LMD
ecoder). LMDecoder saves language-specific audio features into memory banks and
can be trained on audio-text paired data which is more easily accessible than vi
deo-text paired data. Therefore, with LMDecoder, we can transform the input spee
ch units into language-specific audio features and translate them into texts by
utilizing the learned rich language knowledge. Finally, by combining general spe
ech knowledge and language-specific knowledge, we can efficiently develop lip re
ading models even for low-resource languages. Through extensive experiments usin
g five languages, English, Spanish, French, Italian, and Portuguese, the effecti
veness of the proposed method is evaluated.
****

Quality-Agnostic Deepfake Detection with Intra-model Collaborative Learning
Binh M. Le, Simon S. Woo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22378-22389
Deepfake has recently raised a plethora of societal concerns over its possible security threats and dissemination of fake information. Much research on deepfake detection has been undertaken. However, detecting low quality as well as simultaneously detecting different qualities of deepfakes still remains a grave challenge. Most SOTA approaches are limited by using a single specific model for detecting certain deepfake video quality type. When constructing multiple models with prior information about video quality, this kind of strategy incurs significant computational cost, as well as model and training data overhead. Further, it cannot be scalable and practical to deploy in real-world settings. In this work, we propose a universal intra-model collaborative learning framework to enable the effective and simultaneous detection of different quality of deepfakes. That is, our approach is the quality-agnostic deepfake detection method, dubbed QAD . In particular, by observing the upper bound of general error expectation, we maximize the dependency between intermediate representations of images from different quality levels via Hilbert-Schmidt Independence Criterion. In addition, an Adversarial Weight Perturbation module is carefully devised to enable the model to be more robust against image corruption while boosting the overall model's performance. Extensive experiments over seven popular deepfake datasets demonstrate the superiority of our QAD model over prior SOTA benchmarks.
****
Object-Centric Multiple Object Tracking
Zixu Zhao, Jiaze Wang, Max Horn, Yizhuo Ding, Tong He, Zechen Bai, Dominik Zietlow, Carl-Johann Simon-Gabriel, Bing Shuai, Zhuowen Tu, Thomas Brox, Bernt Schiele, Yanwei Fu, Francesco Locatello, Zheng Zhang, Tianjun Xiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16601-16611
Unsupervised object-centric learning methods allow the partitioning of scenes into entities without additional localization information and are excellent candidates for reducing the annotation burden of multiple-object tracking (MOT) pipelines. Unfortunately, they lack two key properties: objects are often split into parts and are not consistently tracked over time. In fact, state-of-the-art models achieve pixel-level accuracy and temporal consistency by relying on supervised object detection with additional ID labels for the association through time. This paper proposes a video object-centric model for MOT. It consists of an index-merge module that adapts the object-centric slots into detection outputs and an object memory module that builds complete object prototypes to handle occlusions. Benefited from object-centric learning, we only require sparse detection labels (0%-6.25%) for object localization and feature binding. Relying on our self-supervised Expectation-Maximization-inspired loss for object association, our approach requires no ID labels. Our experiments significantly narrow the gap between the existing object-centric model and the fully supervised state-of-the-art and outperform several unsupervised trackers that also do not require ID labels.
****
Point-TTA: Test-Time Adaptation for Point Cloud Registration Using Multitask Meta-Auxiliary Learning
Ahmed Hatem, Yiming Qian, Yang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16494-16504
We present Point-TTA, a novel test-time adaptation framework for point cloud registration (PCR) that improves the generalization and the performance of registration models. While learning-based approaches have achieved impressive progress, generalization to unknown testing environments remains a major challenge due to the variations in 3D scans. Existing methods typically train a generic model and the same trained model is applied on each instance during testing. This could be sub-optimal since it is difficult for the same model to handle all the variations during testing. In this paper, we propose a test-time adaptation approach for PCR. Our model can adapt to unseen distributions at test-time without requiring any prior knowledge of the test data. Concretely, we design three self-supervi

sed auxiliary tasks that are optimized jointly with the primary PCR task. Given a test instance, we adapt our model using these auxiliary tasks and the updated model is used to perform the inference. During training, our model is trained using a meta-auxiliary learning approach, such that the adapted model via auxiliary tasks improves the accuracy of the primary task. Experimental results demonstrate the effectiveness of our approach in improving generalization of point cloud registration and outperforming other state-of-the-art approaches.
****
HopFIR: Hop-wise GraphFormer with Intragroup Joint Refinement for 3D Human Pose Estimation
Kai Zhai, Qiang Nie, Bo Ouyang, Xiang Li, Shanlin Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14985-14995
2D-to-3D human pose lifting is fundamental for 3D human pose estimation (HPE), for which graph convolutional networks (GCNs) have proven inherently suitable for modeling the human skeletal topology. However, the current GCN-based 3D HPE methods update the node features by aggregating their neighbors' information without considering the interaction of joints in different joint synergies. Although some studies have proposed importing limb information to learn the movement patterns, the latent synergies among joints, such as maintaining balance are seldom investigated. We propose the Hop-wise GraphFormer with Intragroup Joint Refinement (HopFIR) architecture to tackle the 3D HPE problem. HopFIR mainly consists of a novel hop-wise GraphFormer (HGF) module and an intragroup joint refinement (IJR) module. The HGF module groups the joints by k-hop neighbors and applies a hop-wise transformer-like attention mechanism to these groups to discover latent joint synergies. The IJR module leverages the prior limb information for peripheral joint refinement. Extensive experimental results show that HopFIR outperforms the SOTA methods by a large margin, with a mean per-joint position error (MPJPE) on the Human3.6M dataset of 32.67 mm. We also demonstrate that the state-of-the-art GCN-based methods can benefit from the proposed hop-wise attention mechanism with a significant improvement in performance: SemGCN and MGCN are improved by 8.9% and 4.5%, respectively.
****
Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning
Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, Ge Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4424-4434
Deep neural networks are susceptible to adversarial examples, posing a significant security risk in critical applications. Adversarial Training (AT) is a well-established technique to enhance adversarial robustness, but it often comes at the cost of decreased generalization ability. This paper proposes Robustness Critical Fine-Tuning (RiFT), a novel approach to enhance generalization without compromising adversarial robustness. The core idea of RiFT is to exploit the redundant capacity for robustness by fine-tuning the adversarially trained model on its non-robust-critical module. To do so, we introduce module robust criticality (MRC), a measure that evaluates the significance of a given module to model robustness under worst-case weight perturbations. Using this measure, we identify the module with the lowest MRC value as the non-robust-critical module and fine-tune its weights to obtain fine-tuned weights. Subsequently, we linearly interpolate between the adversarially trained weights and fine-tuned weights to derive the optimal fine-tuned model weights. We demonstrate the efficacy of RiFT on ResNet18, ResNet34, and WideResNet34-10 models trained on CIFAR10, CIFAR100, and Tiny-ImageNet datasets. Our experiments show that RiFT can significantly improve both generalization and out-of-distribution robust- ness by around 1.5% while maintaining or even slightly enhancing adversarial robustness. Code is available at https://github.com/Immortalise/RiFT .
****
Minimal Solutions to Generalized Three-View Relative Pose Problem
Yaqing Ding, Chiang-Heng Chien, Viktor Larsson, Karl Åström, Benjamin Kimia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8156-8164
For a generalized (or non-central) camera model, the minimal problem for two vie

ws of six points has efficient solvers. However, minimal problems of three views with four points and three views of six lines have not yet been explored and solved, despite the efforts from the computer vision community. This paper develops the formulations of these two minimal problems and shows how state-of-the-art GPU implementations of Homotopy Continuation solver can be used effectively. The proposed methods are evaluated on both synthetic and real datasets, demonstrating that they are fast, accurate and that they improve on structure from motion estimations, when employed in an hypothesis and test setting.
****
Trajectory Unified Transformer for Pedestrian Trajectory Prediction
Liushuai Shi, Le Wang, Sanping Zhou, Gang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9675-9684
Pedestrian trajectory prediction is an essentially connecting link to understanding human behavior. Recent works achieve state-of-the-art performance gained from the hand-designed post-processing, e.g., clustering. However, this post-processing suffers from expensive inference time and neglects the probability of the predicted trajectory disturbing downstream safety decisions. In this paper, we present Trajectory Unified TRansformer, called TUTR, which unifies the trajectory prediction components, social interaction and multimodal trajectory prediction, into a transformer encoder-decoder architecture to effectively remove the need for post-processing. Specifically, TUTR parses the relationships across various motion modes by an explicit global prediction and an implicit mode-level transformer encoder. Then, TUTR attends to the social interactions with neighbors by a social-level transformer decoder. Finally, a dual prediction forecasts diverse trajectories and corresponding probabilities in parallel without post-processing. TUTR achieves state-of-the-art accuracy performance and about 10x - 40x inference speed improvements compared with previous well-tuning state-of-the-art methods using post-processing.
****
Understanding the Feature Norm for Out-of-Distribution Detection
Jaewoo Park, Jacky Chen Long Chai, Jaeho Yoon, Andrew Beng Jin Teoh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1557-1567
A neural network trained on a classification dataset often exhibits a higher vector norm of hidden layer features for in-distribution (ID) samples, while producing relatively lower norm values on unseen instances from out-of-distribution (OOD). Despite this intriguing phenomenon being utilized in many applications, the underlying cause has not been thoroughly investigated. In this study, we demystify this very phenomenon by scrutinizing the discriminative structures concealed in the intermediate layers of a neural network. Our analysis leads to the following discoveries: (1) The feature norm is a confidence value of a classifier hidden in the network layer, specifically its maximum logit. Hence, the feature norm distinguishes OOD from ID in the same manner that a classifier confidence does. (2) The feature norm is class-agnostic, thus it can detect OOD samples across diverse discriminative models. (3) The conventional feature norm fails to capture the deactivation tendency of hidden layer neurons, which may lead to misidentification of ID samples as OOD instances. To resolve this drawback, we propose a novel negative-aware norm (NAN) that can capture both the activation and deactivation tendencies of hidden layer neurons. We conduct extensive experiments on NAN, demonstrating its efficacy and compatibility with existing OOD detectors, as well as its capability in label-free environments.
****
MHEntropy: Entropy Meets Multiple Hypotheses for Pose and Shape Recovery
Rongyu Chen, Linlin Yang, Angela Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14840-14849
For monocular RGB-based 3D pose and shape estimation, multiple solutions are often feasible due to factors like occlusion and truncation. This work presents a multi-hypothesis probabilistic framework by optimizing the Kullback-Leibler divergence (KLD) between the data and model distribution. Our formulation reveals a connection between the pose entropy and diversity in the multiple hypotheses that

has been neglected by previous works. For a comprehensive evaluation, besides the best hypothesis (BH) metric, we factor in visibility for evaluating diversity. Additionally, our framework is label-friendly, in that it can be learned from only partial 2D keypoints, e.g., those that are visible. Experiments on both ambiguous and real-world benchmarks demonstrate that our method outperforms other state-of-the-art multi-hypothesis methods in a comprehensive evaluation. The project page is at https://gloryyrolg.github.io/MHEntropy.
****

uSplit: Image Decomposition for Fluorescence Microscopy

Ashesh Ashesh, Alexander Krull, Moises Di Sante, Francesco Pasqualini, Florian Jug; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21219-21229

We present mSplit, a dedicated approach for trained image decomposition in the context of fluorescence microscopy images. We find that best results using regular deep architectures are achieved when large image patches are used during training, making memory consumption the limiting factor to further improving performance. We therefore introduce lateral contextualization (LC), a novel meta-architecture that enables the memory efficient incorporation of large image-context, which we observe is a key ingredient to solving the image decomposition task at hand. We integrate LC with U-Nets, Hierarchical AEs, and Hierarchical VAEs, for which we formulate a modified ELBO loss. Additionally, LC enables training deeper hierarchical models than otherwise possible and, interestingly, helps to reduce tiling artefacts that are inherently impossible to avoid when using tiled VAE predictions. We apply mSplit to five decomposition tasks, one on a synthetic dataset, four others derived from real microscopy data. Our method consistently achieves best results (average improvements to the best baseline of 2.25 dB PSNR), while simultaneously requiring considerably less GPU memory. Our code and datasets can be found at https://github.com/juglab/uSplit.
****

Modeling the Relative Visual Tempo for Self-supervised Skeleton-based Action Recognition

Yisheng Zhu, Hu Han, Zhengtao Yu, Guangcan Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13913-13922

Visual tempo characterizes the dynamics and the temporal evolution, which helps describe actions. Recent approaches directly perform visual tempo prediction on skeleton sequences, which may suffer from insufficient feature representation issue. In this paper, we observe that relative visual tempo is more in line with human intuition, and thus providing more effective supervision signals. Based on this, we propose a novel Relative Visual Tempo Contrastive Learning framework for skeleton action Representation (RVTCLR). Specifically, we design a Relative Visual Tempo Learning (RVTL) task to explore the motion information in intra-video clips, and an Appearance-Consistency (AC) task to learn appearance information simultaneously, resulting in more representative spatiotemporal features. Furthermore, skeleton sequence data is much sparser than RGB data, making the network learn shortcuts, and overfit to low-level information such as skeleton scales. To learn high-order semantics, we further design a new Distribution-Consistency (DC) branch, containing three components: Skeleton-specific Data Augmentation (SDA), Fine-grained Skeleton Encoding Module (FSEM), and Distribution-aware Diversity (DD) Loss. We term our entire method (RVTCLR with DC) as RVTCLR+. Extensive experiments on NTU RGB+D 60 and NTU RGB+D 120 datasets demonstrate that our RVTCLR+ can achieve competitive results over the state-of-the-art methods. Code is available at https://github.com/Zhuysheng/RVTCLR.
****

LightGlue: Local Feature Matching at Light Speed

Philipp Lindenberger, Paul-Edouard Sarlin, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17627-17638

We introduce LightGlue, a deep neural network that learns to match local features across images. We revisit multiple design decisions of SuperGlue, the state of the art in sparse matching, and derive simple but effective improvements. Cumulatively, they make LightGlue more efficient -- in terms of both memory and compu

tation, more accurate, and much easier to train. One key property is that LightGlue is adaptive to the difficulty of the problem: the inference is much faster on image pairs that are intuitively easy to match, for example because of a larger visual overlap or limited appearance change. This opens up exciting prospects for deploying deep matchers in latency-sensitive applications like 3D reconstruction. The code and trained models are publicly available at github.com/cvg/LightGlue.
****
Masked Autoencoders are Efficient Class Incremental Learners
Jiang-Tian Zhai, Xialei Liu, Andrew D. Bagdanov, Ke Li, Ming-Ming Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19104-19113
Class Incremental Learning (CIL) aims to sequentially learn new classes while avoiding catastrophic forgetting of previous knowledge. We propose to use Masked Autoencoders (MAEs) as efficient learners for CIL. MAEs were originally designed to learn useful representations through reconstructive unsupervised learning, and they can be easily integrated with a supervised loss for classification. Moreover, MAEs can reliably reconstruct original input images from randomly selected patches, which we use to store exemplars from past tasks more efficiently for CIL. We also propose a bilateral MAE framework to learn from image-level and embedding-level fusion, which produces better-quality reconstructed images and more stable representations. Our experiments confirm that our approach performs better than the state-of-the-art on CIFAR-100, ImageNet-Subset, and ImageNet-Full. The code is available at https://github.com/scok30/MAE-CIL.
****
Knowledge Proxy Intervention for Deconfounded Video Question Answering
Jiangtong Li, Li Niu, Liqing Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2782-2793
Recently, Video Question-Answering (VideoQA) has drawn more and more attention from both industry and research community. Despite all the success achieved by recent works, dataset bias always harmfully misleads current methods focusing on spurious correlations in training data. To analyze the effects of dataset bias, we frame the VideoQA pipeline into a causal graph, which shows the causalities among video, question, aligned feature between video and question, answer, and underlying confounder. Through the causal graph, we prove that the confounder and the backdoor path lead to spurious causality. To tackle the challenge that the confounder in VideoQA is unobserved and non-enumerable in general, we propose a model-agnostic framework called Knowledge Proxy Intervention (KPI), which introduces an extra knowledge proxy variable in the causal graph to cut the backdoor path and remove the confounder. Our KPI framework exploits the front-door adjustment, which requires no prior knowledge about the confounder. The effectiveness of our KPI framework is corroborated by three baseline methods on five benchmark datasets, including MSVD-QA, MSRVTT-QA, TGIF-QA, NExT-QA, and Causal-VidQA.
****
Towards Semi-supervised Learning with Non-random Missing Labels
Yue Duan, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, Yinghuan Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16121-16131
Semi-supervised learning (SSL) tackles the label missing problem by enabling the effective usage of unlabeled data. While existing SSL methods focus on the traditional setting, a practical and challenging scenario called label Missing Not At Random (MNAR) is usually ignored. In MNAR, the labeled and unlabeled data fall into different class distributions resulting in biased label imputation, which deteriorates the performance of SSL models. In this work, class transition tracking based Pseudo-Rectifying Guidance (PRG) is devised for MNAR. We explore the class-level guidance information obtained by the Markov random walk, which is mod

eled on a dynamically created graph built over the class tracking matrix. PRG un
ifies the history information of each class transition caused by the pseudo-rect
ifying procedure to activate the model's enthusiasm for neglected classes, so as
 the quality of pseudo-labels on both popular classes and rare classes in MNAR c
ould be improved. We show the superior performance of PRG across a variety of MN
AR scenarios, outperforming the latest SSL approaches combining bias removal sol
utions by a large margin. Code and model weights are available at https://github
.com/NJUyued/PRG4SSL-MNAR.
****
DetZero: Rethinking Offboard 3D Object Detection with Long-term Sequential Point
 Clouds
Tao Ma, Xuemeng Yang, Hongbin Zhou, Xin Li, Botian Shi, Junjie Liu, Yuchen Yang,
 Zhizheng Liu, Liang He, Yu Qiao, Yikang Li, Hongsheng Li; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6736-6747
Existing offboard 3D detectors always follow a modular pipeline design to take a
dvantage of unlimited sequential point clouds. We have found that the full poten
tial of offboard 3D detectors is not explored mainly due to two reasons: (1) the
 onboard multi-object tracker cannot generate sufficient complete object traject
ories, and (2) the motion state of objects poses an inevitable challenge for the
 object-centric refining stage in leveraging the long-term temporal context repr
esentation. To tackle these problems, we propose a novel paradigm of offboard 3D
 object detection, named DetZero. Concretely, an offline tracker coupled with a
multi-frame detector is proposed to focus on the completeness of generated objec
t tracks. An attention-mechanism refining module is proposed to strengthen conte
xtual information interaction across long-term sequential point clouds for objec
t refining with decomposed regression methods. Extensive experiments on Waymo Op
en Dataset show our DetZero outperforms all state-of-the-art onboard and offboar
d 3D detection methods. Notably, DetZero ranks 1st place on Waymo 3D object dete
ction leaderboard with 85.15 mAPH (L2) detection performance. Further experiment
s validate the application of taking the place of human labels with such high-qu
ality results. Our empirical study leads to rethinking conventions and interesti
ng findings that can guide future research on offboard 3D object detection.
****
ImbSAM: A Closer Look at Sharpness-Aware Minimization in Class-Imbalanced Recogn
ition
Yixuan Zhou, Yi Qu, Xing Xu, Hengtao Shen; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2023, pp. 11345-11355
Class imbalance is a common challenge in real-world recognition tasks, where the
 majority of classes have few samples, also known as tail classes. We address th
is challenge with the perspective of generalization and empirically find that th
e promising Sharpness-Aware Minimization (SAM) fails to address generalization i
ssues under the class-imbalanced setting. Through investigating this specific ty
pe of task, we identify that its generalization bottleneck primarily lies in the
 severe overfitting for tail classes with limited training data. To overcome thi
s bottleneck, we leverage class priors to restrict the generalization scope of t
he class-agnostic SAM and propose a class-aware smoothness optimization algorith
m named Imbalanced-SAM (ImbSAM). With the guidance of class priors, our ImbSAM s
pecifically improves generalization targeting tail classes. We also verify the e
fficacy of ImbSAM on two prototypical applications of class-imbalanced recogniti
on: long-tailed classification and semi-supervised anomaly detection, where our
ImbSAM demonstrates remarkable performance improvements for tail classes and ano
maly. Our code implementation is available at https://github.com/cool-xuan/Imbal
anced_SAM.
****
Learning from Noisy Data for Semi-Supervised 3D Object Detection
Zehui Chen, Zhenyu Li, Shuo Wang, Dengpan Fu, Feng Zhao; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6929-6939
Pseudo-Labeling (PL) is a critical approach in semi-supervised 3D object detecti
on (SSOD). In PL, delicately selected pseudo-labels, generated by the teacher mo
del, are provided for the student model to supervise the semi-supervised detecti

on framework. However, such a paradigm may introduce misclassified labels or loose localized box predictions, resulting in a sub-optimal solution of detection performance. In this paper, we take PL from a noisy learning perspective: instead of directly applying vanilla pseudo-labels, we design a noise-resistant instance supervision module for better generalization. Specifically, we soften the classification targets by considering both the quality of pseudo labels and the network learning ability, and convert the regression task into a probabilistic modeling problem. Besides, considering that self-supervised learning works in the absence of labels, we incorporate dense pixel-wise feature consistency constraints to eliminate the negative impact of noisy labels. To this end, we propose NoiseDet, a simple yet effective framework for semi-supervised 3D object detection. Extensive experiments on competitive ONCE and Waymo benchmarks demonstrate that our method outperforms current semi-supervised approaches by a large margin. Notably, our NoiseDet achieves state-of-the-art performance under various dataset scales on ONCE dataset. For example, NoiseDet improves its NoiseyStudent baseline from 55.5 mAP to 58.0 mAP, and further reaches 60.2 mAP with enhanced pseudo-label generation. Code will be available at https://github.com/zehuichen123/NoiseDet.
****
NeRFrac: Neural Radiance Fields through Refractive Surface
Yifan Zhan, Shohei Nobuhara, Ko Nishino, Yinqiang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18402-18412
Neural Radiance Fields (NeRF) is a popular neural expression for novel view synthesis. By querying spatial points and view directions, a multilayer perceptron (MLP) can be trained to output the volume density and radiance at each point, which lets us render novel views of the scene. The original NeRF and its recent variants, however, target opaque scenes dominated by diffuse reflection surfaces and cannot handle complex refractive surfaces well. We introduce NeRFrac to realize neural novel view synthesis of scenes captured through refractive surfaces, typically water surfaces. For each queried ray, an MLP-based Refractive Field is trained to estimate the distance from the ray origin to the refractive surface. A refracted ray at each intersection point is then computed by Snell's Law, given the input ray and the approximated local normal. Points of the scene are sampled along the refracted ray and are sent to a Radiance Field for further radiance estimation. We show that from a sparse set of images, our model achieves accurate novel view synthesis of the scene underneath the refractive surface and simultaneously reconstructs the refractive surface. We evaluate the effectiveness of our method with synthetic and real scenes seen through water surfaces. Experimental results demonstrate the accuracy of NeRFrac for modeling scenes seen through wavy refractive surfaces.
****
MonoDETR: Depth-guided Transformer for Monocular 3D Object Detection
Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, Peng Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9155-9166
Monocular 3D object detection has long been a challenging task in autonomous driving. Most existing methods follow conventional 2D detectors to first localize object centers, and then predict 3D attributes by neighboring features. However, only using local visual features is insufficient to understand the scene-level 3D spatial structures and ignores the long-range inter-object depth relations. In this paper, we introduce the first DETR framework for Monocular DEtection with a depth-guided TRansformer, named MonoDETR. We modify the vanilla transformer to be depth-aware and guide the whole detection process by contextual depth cues. Specifically, concurrent to the visual encoder that captures object appearances, we introduce to predict a foreground depth map, and specialize a depth encoder to extract non-local depth embeddings. Then, we formulate 3D object candidates as learnable queries and propose a depth-guided decoder to conduct object-scene depth interactions. In this way, each object query estimates its 3D attributes adaptively from the depth-guided regions on the image and is no longer constrained to local visual features. On KITTI benchmark with monocular images as input, Mo

noDETR achieves state-of-the-art performance and requires no extra dense depth annotations. Besides, our depth-guided modules can also be plug-and-play to enhance multi-view 3D object detectors on nuScenes dataset, demonstrating our superior generalization capacity. Code is available at https://github.com/ZrrSkywalker/MonoDETR.
****
Towards Authentic Face Restoration with Iterative Diffusion Models and Beyond
Yang Zhao, Tingbo Hou, Yu-Chuan Su, Xuhui Jia, Yandong Li, Matthias Grundmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7312-7322
An authentic face restoration system is becoming increasingly demanding in many computer vision applications, e.g., image enhancement, video communication, and taking portrait. Most of the advanced face restoration models can recover high-quality faces from low-quality ones but usually fail to faithfully generate realistic and high-frequency details that are favored by users. To achieve authentic restoration, we propose IDM, an Iteratively learned face restoration system based on denoising Diffusion Models (DDMs). We define the criterion of an authentic face restoration system, and argue that denoising diffusion models are naturally endowed with this property from two aspects: intrinsic iterative refinement and extrinsic iterative enhancement. Intrinsic learning can preserve the content well and gradually refine the high-quality details, while extrinsic enhancement helps clean the data and improve the restoration task one step further. We demonstrate superior performance on blind face restoration tasks. Beyond restoration, we find the authentically cleaned data by the proposed restoration system is also helpful to image generation tasks in terms of training stabilization and sample quality. Without modifying the baseline models, we achieve better quality than state-of-the-art on FFHQ and ImageNet generation using either GANs or diffusion models.
****
LivelySpeaker: Towards Semantic-Aware Co-Speech Gesture Generation
Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, Shenghua Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20807-20817
Gestures are non-verbal but important behaviors accompanying people's speech. While previous methods are able to generate speech rhythm-synchronized gestures, the semantic context of the speech is generally lacking in the gesticulations. Although semantic gestures do not occur very regularly in human speech, they are indeed the key for the audience to understand the speech context in a more immersive environment. Hence, we introduce LivelySpeaker, a framework that realizes semantics-aware co-speech gesture generation and offers several control handles. Specifically, the script-based gesture generation leverages the pre-trained CLIP text embeddings as the guidance for generating gestures that are highly semantically aligned with the script. Then, we devise a simple but effective diffusion-based gesture generation backbone simply using pure MLPs, that is conditioned on only audio signals and learns to gesticulate with realistic motions. We utilize such powerful prior to rhyme the script-guided gestures with the audio signals, notably in a zero-shot setting. Our novel two-stage generation framework also enables several applications, such as changing the gesticulation style, editing the co-speech gestures via textual prompting, and controlling the semantic awareness and rhythm alignment with guided diffusion. Extensive experiments demonstrate the advantages of the proposed framework over competing methods. In addition, our core diffusion-based generative model also achieves state-of-the-art performance on two benchmarks. The code and model will be released to facilitate future research.
****
Contrastive Feature Masking Open-Vocabulary Vision Transformer
Dahun Kim, Anelia Angelova, Weicheng Kuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15602-15612
We present Contrastive Feature Masking Vision Transformer (CFM-ViT) - an image-text pretraining methodology that achieves simultaneous learning of image- and re

gion level representation for open-vocabulary object detection (OVD). Our approach combines the masked autoencoder (MAE) objective into the contrastive learning objective to improve the representation for localization tasks. Unlike standard MAE, we perform reconstruction in the joint image-text embedding space, rather than the pixel space as is customary with the classical MAE method, which causes the model to better learn region-level semantics. Moreover, we introduce Positional Embedding Dropout (PED) to address scale variation between image-text pretraining and detection finetuning by randomly dropping out the positional embeddings during pretraining. PED improves detection performance and enables the use of a frozen ViT backbone as a region classifier, preventing the forgetting of open-vocabulary knowledge during detection finetuning. On LVIS open-vocabulary detection benchmark, CFM-ViT achieves a state-of-the-art 33.9 APr, surpassing the best approach by 7.6 points and achieves better zero-shot detection transfer. Finally, CFM-ViT acquires strong image-level representation, outperforming the state of the art on 8 out of 12 metrics on zero-shot image-text retrieval benchmarks.
****
Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment
Qiang Chen, Xiaokang Chen, Jian Wang, Shan Zhang, Kun Yao, Haocheng Feng, Junyu Han, Errui Ding, Gang Zeng, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6633-6642
Detection transformer (DETR) relies on one-to-one assignment, assigning one ground-truth object to one prediction, for end-to-end detection without NMS post-processing. It is known that one-to-many assignment, assigning one ground-truth object to multiple predictions, succeeds in detection methods such as Faster R-CNN and FCOS. While the naive one-to-many assignment does not work for DETR, and it remains challenging to apply one-to-many assignment for DETR training.

In this paper, we introduce Group DETR, a simple yet efficient DETR training approach that introduces a group-wise way for one-to-many assignment. This approach involves using multiple groups of object queries, conducting one-to-one assignment within each group, and performing decoder self-attention separately. It resembles data augmentation with automatically-learned object query augmentation. It is also equivalent to simultaneously training parameter-sharing networks of the same architecture, introducing more supervision and thus improving DETR training. The inference process is the same as DETR trained normally and only needs one group of queries without any architecture modification. Group DETR is versatile and is applicable to various DETR variants. The experiments show that Group DETR significantly speeds up the training convergence and improves the performance of various DETR-based models. Code will be available at https://github.com/Atten4Vis/GroupDETR.
****
Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models
Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, Yang You; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19125-19136
Continual learning (CL) can help pre-trained vision-language models efficiently adapt to new or under-trained data distributions without re-training. Nevertheless, during the continual training of the Contrastive Language-Image Pre-training (CLIP) model, we observe that the model's zero-shot transfer ability significantly degrades due to catastrophic forgetting. Existing CL methods can mitigate forgetting by replaying previous data. However, since the CLIP dataset is private, replay methods cannot access the pre-training dataset. In addition, replaying data of previously learned downstream tasks can enhance their performance but comes at the cost of sacrificing zero-shot performance. To address this challenge, we propose a novel method ZSCL to prevent zero-shot transfer degradation in the continual learning of vision-language models in both feature and parameter space. In the feature space, a reference dataset is introduced for distillation between the current and initial models. The reference dataset should have semantic diversity but no need to be labeled, seen in pre-training, or matched image-text p

airs. In parameter space, we prevent a large parameter shift by averaging weights during the training. We propose a more challenging Multi-domain Task Incremental Learning (MTIL) benchmark to evaluate different methods, where tasks are from various domains instead of class-separated in a single dataset. Our method outperforms other methods in the traditional class-incremental learning setting and the MTIL by 9.7% average score. Our code locates at https://github.com/Thunderbeee/ZSCL.
****
Personalized Image Generation for Color Vision Deficiency Population
Shuyi Jiang, Daochang Liu, Dingquan Li, Chang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22571-22580
Approximately, 350 million people, a proportion of 8%, suffer from color vision deficiency (CVD). While image generation algorithms have been highly successful in synthesizing high-quality images, CVD populations are unintentionally excluded from target users and have difficulties understanding the generated images as normal viewers do. Although a straightforward baseline can be formed by combining generation models and recolor compensation methods as the post-processing, the CVD friendliness of the result images is still limited since the input image content of recolor methods is not CVD-oriented and will keep fixed during the recolor compensation process. Besides, the CVD populations can't be fully served since the varying degrees of CVD are often neglected in recoloring methods. Instead, we propose a personalized CVD-friendly image generation algorithm with two key characteristics: (i) generating CVD-oriented images end-to-end; (ii) generating continuous personalized images for people with various CVD types and degrees through disentangling the color representation based on a triple-latent structure. Quantitative experiments and the user study indicate our proposed image generation model can generate practical and compelling results compared to the normal generation model and combination baselines on several datasets.
****
EGC: Image Generation and Classification via a Diffusion Energy-Based Model
Qiushan Guo, Chuofan Ma, Yi Jiang, Zehuan Yuan, Yizhou Yu, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22952-22962
Learning image classification and image generation using the same set of network parameters presents a formidable challenge. Recent advanced approaches perform well in one task often exhibit poor performance in the other. This work introduces an energy-based classifier and generator, namely EGC, which can achieve superior performance in both tasks using a single neural network. Unlike conventional classifiers that produce a label given an image (i.e., a conditional distribution p(y|x)), the forward pass in EGC is a classification model that yields a joint distribution p(x,y), enabling a diffusion model in its backward pass by marginalizing out the label y to estimate the score function. Furthermore, EGC can be adapted for unsupervised learning by considering the label as latent variables. EGC achieves competitive generation results compared with state-of-the-art approaches on ImageNet-1k, CelebA-HQ and LSUN Church, while achieving superior classification accuracy and robustness against adversarial attacks on CIFAR-10. This work marks the inaugural success in mastering both domains using a unified network parameter set.
We believe that EGC bridges the gap between discriminative and generative learning.
****
OccFormer: Dual-path Transformer for Vision-based 3D Semantic Occupancy Prediction
Yunpeng Zhang, Zheng Zhu, Dalong Du; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9433-9443
The vision-based perception for autonomous driving has undergone a transformation from the bird-eye-view (BEV) representations to the 3D semantic occupancy. Compared with the BEV planes, the 3D semantic occupancy further provides structural information along the vertical direction. This paper presents OccFormer, a dual-path transformer network to effectively process the 3D volume for semantic occu

pancy prediction. OccFormer achieves a long-range, dynamic, and efficient encoding of the camera-generated 3D voxel features. It is obtained by decomposing the heavy 3D processing into the local and global transformer pathways along the horizontal plane. For the occupancy decoder, we adapt the vanilla Mask2Former for 3D semantic occupancy by proposing preserve-pooling and classguided sampling, which notably mitigate the sparsity and class imbalance. Experimental results demonstrate that OccFormer significantly outperforms existing methods for semantic scene completion on SemanticKITTI dataset and for LiDAR semantic segmentation on nuScenes dataset. Code is available at https://github.com/zhangyp15/OccFormer.
****
Probabilistic Triangulation for Uncalibrated Multi-View 3D Human Pose Estimation
Boyuan Jiang, Lei Hu, Shihong Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14850-14860
3D human pose estimation has been a long-standing challenge in computer vision and graphics, where multi-view methods have significantly progressed but are limited by the tedious calibration processes. Existing multi-view methods are restricted to fixed camera pose and therefore lack generalization ability. This paper presents a novel Probabilistic Triangulation module that can be embedded in a calibrated 3D human pose estimation method, generalizing it to uncalibration scenes. The key idea is to use a probability distribution to model the camera pose and iteratively update the distribution from 2D features instead of using camera pose. Specifically, We maintain a camera pose distribution and then iteratively update this distribution by computing the posterior probability of the camera pose through Monte Carlo sampling. This way, the gradients can be directly back-propagated from the 3D pose estimation to the 2D heatmap, enabling end-to-end training. Extensive experiments on Human3.6M and CMU Panoptic demonstrate that our method outperforms other uncalibration methods and achieves comparable results with state-of-the-art calibration methods. Thus, our method achieves a trade-off between estimation accuracy and generalizability.
****
Joint Metrics Matter: A Better Standard for Trajectory Forecasting
Erica Weng, Hana Hoshino, Deva Ramanan, Kris Kitani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20315-20326
Multi-modal trajectory forecasting methods commonly evaluate using single-agent metrics (marginal metrics), such as minimum Average Displacement Error (ADE) and Final Displacement Error (FDE), which fail to capture joint performance of multiple interacting agents. Only focusing on marginal metrics can lead to unnatural predictions, such as colliding trajectories or diverging trajectories for people who are clearly walking together as a group. Consequently, methods optimized for marginal metrics lead to overly-optimistic estimations of performance, which is detrimental to progress in trajectory forecasting research.
In response to the limitations of marginal metrics, we present the first comprehensive evaluation of state-of-the-art (SOTA) trajectory forecasting methods with respect to multi-agent metrics (joint metrics): JADE, JFDE, and collision rate. We demonstrate the importance of joint metrics as opposed to marginal metrics with quantitative evidence and qualitative examples drawn from the ETH / UCY and Stanford Drone datasets. We introduce a new loss function incorporating joint metrics that, when applied to a SOTA trajectory forecasting method, achieves SOTA performance with respect to JADE and JFDE, achieving a 7% improvement over the previous SOTA on the ETH / UCY datasets. Our results also indicate that optimizing for joint metrics naturally leads to an improvement in interaction modeling, as evidenced by a 16% decrease in mean collision rate on the ETH / UCY datasets with respect to the previous SOTA. Code is available at https://github.com/ericaweng/joint-metrics-matter.
****
TORE: Token Reduction for Efficient Human Mesh Recovery with Transformer
Zhiyang Dou, Qingxuan Wu, Cheng Lin, Zeyu Cao, Qiangqiang Wu, Weilin Wan, Taku Komura, Wenping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15143-15155
In this paper, we introduce a set of simple yet effective TOken REduction (TORE)

strategies for Transformer-based Human Mesh Recovery from monocular images. Current SOTA performance is achieved by Transformer-based structures. However, they suffer from high model complexity and computation cost caused by redundant tokens. We propose token reduction strategies based on two important aspects, i.e., the 3D geometry structure and 2D image feature, where we hierarchically recover the mesh geometry with priors from body structure and conduct token clustering to pass fewer but more discriminative image feature tokens to the Transformer. Our method massively reduces the number of tokens involved in high-complexity interactions in the Transformer. This leads to a significantly reduced computational cost while still achieving competitive or even higher accuracy in shape recovery. Extensive experiments across a wide range of benchmarks validate the superior effectiveness of the proposed method. We further demonstrate the generalizability of our method on hand mesh recovery. Visit our project page at https://frank-zy-dou.github.io/projects/Tore/index.html.
****
Test Time Adaptation for Blind Image Quality Assessment
Subhadeep Roy, Shankhanil Mitra, Soma Biswas, Rajiv Soundararajan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16742-16751
While the design of blind image quality assessment (IQA) algorithms has improved significantly, the distribution shift between the training and testing scenarios often leads to a poor performance of these methods at inference time. This motivates the study of test time adaptation (TTA) techniques to improve their performance at inference time. Existing auxiliary tasks and loss functions used for TTA may not be relevant for quality-aware adaptation of the pre-trained model. In this work, we introduce two novel quality-relevant auxiliary tasks at the batch and sample levels to enable TTA for blind IQA. In particular, we introduce a group contrastive loss at the batch level and a relative rank loss at the sample level to make the model quality aware and adapt to the target data. Our experiments reveal that even using a small batch of images from the test distribution helps achieve significant improvement in performance by updating the batch normalization statistics of the source model.
****
GeT: Generative Target Structure Debiasing for Domain Adaptation
Can Zhang, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23577-23588
Domain adaptation (DA) aims to transfer knowledge from a fully labeled source to a scarcely labeled or totally unlabeled target under domain shift. Recently, semi-supervised learning-based (SSL) techniques that leverage pseudo labeling have been increasingly used in DA. Despite the competitive performance, these pseudo labeling methods rely heavily on the source domain to generate pseudo labels for the target domain and therefore still suffer considerably from source data bias. Moreover, class distribution bias in the target domain is also often ignored in the pseudo label generation and thus leading to further deterioration of performance. In this paper, we propose GeT that learns a non-bias target embedding distribution with high quality pseudo labels. Specifically, we formulate an online target generative classifier to induce the target distribution into distinctive Gaussian components weighted by their class priors to mitigate source data bias and enhance target class discriminability. We further propose a structure similarity regularization framework to alleviate target class distribution bias and further improve target class discriminability. Experimental results show that our proposed GeT is effective and achieves consistent improvements under various DA settings with and without class distribution bias. Our code is available at: https://lulusindazc.github.io/getproject/.
****
D3G: Exploring Gaussian Prior for Temporal Sentence Grounding with Glance Annotation
Hanjun Li, Xiujun Shu, Sunan He, Ruizhi Qiao, Wei Wen, Taian Guo, Bei Gan, Xing Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13734-13746

Temporal sentence grounding (TSG) aims to locate a specific moment from an untrimmed video with a given natural language query. Recently, weakly supervised methods still have a large performance gap compared to fully supervised ones, while the latter requires laborious timestamp annotations. In this study, we aim to reduce the annotation cost yet keep competitive performance for TSG task compared to fully supervised ones. To achieve this goal, we investigate a recently proposed glance-supervised temporal sentence grounding task, which requires only single frame annotation (referred to as glance annotation) for each query. Under this setup, we propose a Dynamic Gaussian prior based Grounding framework with Glance annotation (D3G), which consists of a Semantic Alignment Group Contrastive Learning module (SA-GCL) and a Dynamic Gaussian prior Adjustment module (DGA). Specifically, SA-GCL samples reliable positive moments from a 2D temporal map via jointly leveraging Gaussian prior and semantic consistency, which contributes to aligning the positive sentence-moment pairs in the joint embedding space. Moreover, to alleviate the annotation bias resulting from glance annotation and model complex queries consisting of multiple events, we propose the DGA module, which adjusts the distribution dynamically to approximate the ground truth of target moments. Extensive experiments on three challenging benchmarks verify the effectiveness of the proposed D3G. It outperforms the state-of-the-art weakly supervised methods by a large margin and narrows the performance gap compared to fully supervised methods. Code is available at https://github.com/solicucu/D3G.
****
GEDepth: Ground Embedding for Monocular Depth Estimation
Xiaodong Yang, Zhuang Ma, Zhiyu Ji, Zhe Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12719-12727
Monocular depth estimation is an ill-posed problem as the same 2D image can be projected from infinite 3D scenes. Although the leading algorithms in this field have reported significant improvement, they are essentially geared to the particular compound of pictorial observations and camera parameters (i.e., intrinsics and extrinsics), strongly limiting their generalizability in real-world scenarios. In order to cope with this difficulty, this paper proposes a novel ground embedding module to decouple camera parameters from pictorial cues, thus promoting the generalization capability. Given camera parameters, our module generates the ground depth, which is stacked with the input image and referenced in the final depth prediction. A ground attention is designed in the module to optimally combine the ground depth with the residual depth. The proposed ground embedding is highly flexible and lightweight, leading to a plug-in module that is amenable to be integrated into various depth estimation networks. Experiments reveal that our approach achieves the state-of-the-art results on popular benchmarks, and more importantly, renders significant improvement on the cross-domain generalization.
****
DETRs with Collaborative Hybrid Assignments Training
Zhuofan Zong, Guanglu Song, Yu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6748-6758
In this paper, we provide the observation that too few queries assigned as positive samples in DETR with one-to-one set matching leads to sparse supervision on the encoder's output which considerably hurt the discriminative feature learning of the encoder and vice visa for attention learning in the decoder. To alleviate this, we present a novel collaborative hybrid assignments training scheme, namely Co-DETR, to learn more efficient and effective DETR-based detectors from versatile label assignment manners. This new training scheme can easily enhance the encoder's learning ability in end-to-end detectors by training the multiple parallel auxiliary heads supervised by one-to-many label assignments such as ATSS and Faster RCNN. In addition, we conduct extra customized positive queries by ext

racting the positive coordinates from these auxiliary heads to improve the training efficiency of positive samples in the decoder. In inference, these auxiliary heads are discarded and thus our method introduces no additional parameters and computational cost to the original detector while requiring no hand-crafted non-maximum suppression (NMS). We conduct extensive experiments to evaluate the effectiveness of the proposed approach on DETR variants, including DAB-DETR, Deformable-DETR, and DINO-Deformable-DETR. The state-of-the-art DINO-Deformable-DETR with Swin-L can be improved from 58.5% to 59.5% AP on COCO val. Surprisingly, incorporated with ViT-L backbone, we achieve 66.0% AP on COCO test-dev and 67.9% AP on LVIS val, outperforming previous methods by clear margins with much fewer model sizes. Codes are available at https://github.com/Sense-X/Co-DETR.
****

Animal3D: A Comprehensive Dataset of 3D Animal Pose and Shape
Jiacong Xu, Yi Zhang, Jiawei Peng, Wufei Ma, Artur Jesslen, Pengliang Ji, Qixin Hu, Jiehua Zhang, Qihao Liu, Jiahao Wang, Wei Ji, Chen Wang, Xiaoding Yuan, Prakhar Kaushik, Guofeng Zhang, Jie Liu, Yushan Xie, Yawen Cui, Alan Yuille, Adam Kortylewski; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9099-9109
Accurately estimating the 3D pose and shape is an essential step towards understanding animal behavior, and can potentially benefit many downstream applications, such as wildlife conservation. However, research in this area is held back by the lack of a comprehensive and diverse dataset with high-quality 3D pose and shape annotations. In this paper, we propose Animal3D, the first comprehensive dataset for mammal animal 3D pose and shape estimation. Animal3D consists of 3379 images collected from 40 mammal species, high-quality annotations of 26 keypoints, and importantly the pose and shape parameters of the SMAL model. All annotations were labeled and checked manually in a multi-stage process to ensure highest quality results. Based on the Animal3D dataset, we benchmark representative shape and pose estimation models at: (1) supervised learning from only the Animal3D data, (2) synthetic to real transfer from synthetically generated images, and (3) fine-tuning human pose and shape estimation models. Our experimental results demonstrate that predicting the 3D shape and pose of animals across species remains a very challenging task, despite significant advances in human pose estimation and animal pose estimation for specific species. Our results further demonstrate that synthetic pre-training is a viable strategy to boost the model performance. Overall, Animal3D opens new directions for facilitating future research in animal 3D pose and shape estimation, and is publicly available.
****

Rethinking Video Frame Interpolation from Shutter Mode Induced Degradation
Xiang Ji, Zhixiang Wang, Zhihang Zhong, Yinqiang Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12259-12268
Image restoration from various motion-related degradations, like blurry effects recorded by a global shutter (GS) and jello effects caused by a rolling shutter (RS), has been extensively studied. It has been recently recognized that such degradations encode temporal information, which can be exploited for video frame interpolation (VFI), a more challenging task than pure restoration. However, these VFI researches are mainly grounded on experiments with synthetic data, rather than real data. More fundamentally, under the same imaging condition, it remains unknown which degradation will be more effective toward VFI. In this paper, we present the first real-world dataset for learning and benchmark degraded video frame interpolation, named RD-VFI, and further explore the performance differences of three types of degradations, including GS blur, RS distortion, and an in-between effect caused by the rolling shutter with global reset (RSGR), thanks to our novel quad-axis imaging system. Moreover, we propose a unified Progressive Mutual Boosting Network (PMBNet) model to interpolate middle frames at arbitrary times for all shutter modes. Its disentanglement strategy and dual-stream correction enable us to adaptively deal with different degradations for VFI. Experimental results demonstrate that our PMBNet is superior to the respective state-of-the-art methods on all shutter modes.
****

Multi-Modal Neural Radiance Field for Monocular Dense SLAM with a Light-Weight T oF Sensor
Xinyang Liu, Yijin Li, Yanbin Teng, Hujun Bao, Guofeng Zhang, Yinda Zhang, Zhaop eng Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1-11
Light-weight time-of-flight (ToF) depth sensors are compact and cost-efficient, and thus widely used on mobile devices for tasks such as autofocus and obstacle detection. However, due to the sparse and noisy depth measurements, these sensor s have rarely been considered for dense geometry reconstruction. In this work, w e present the first dense SLAM system with a monocular camera and a light-weight ToF sensor. Specifically, we propose a multi-modal implicit scene representatio n that supports rendering both the signals from the RGB camera and light-weight ToF sensor which drives the optimization by comparing with the raw sensor inputs . Moreover, in order to guarantee successful pose tracking and reconstruction, w e exploit a predicted depth as an intermediate supervision and develop a coarse-to-fine optimization strategy for efficient learning of the implicit representat ion. At last, the temporal information is explicitly exploited to deal with the noisy signals from light-weight ToF sensors to improve the accuracy and robustne ss of the system. Experiments demonstrate that our system well exploits the sign als of light-weight ToF sensors and achieves competitive results both on camera tracking and dense scene reconstruction. Project page: https://zju3dv.github.io/ tof_slam/.
****
MonoNeRD: NeRF-like Representations for Monocular 3D Object Detection
Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, Deng Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2023, pp. 6814-6824
In the field of monocular 3D detection, it is common practice to utilize scene g eometric clues to enhance the detector's performance. However, many existing wor ks adopt these clues explicitly such as estimating a depth map and back-projecti ng it into 3D space. This explicit methodology induces sparsity in 3D representa tions due to the increased dimensionality from 2D to 3D, and leads to substantia l information loss, especially for distant and occluded objects. To alleviate th is issue, we propose MonoNeRD, a novel detection framework that can infer dense 3D geometry and occupancy. Specifically, we model scenes with Signed Distance Fu nctions (SDF), facilitating the production of dense 3D representations. We treat these representations as Neural Radiance Fields (NeRF) and then employ volume r endering to recover RGB images and depth maps. To the best of our knowledge, thi s work is the first to introduce volume rendering for M3D, and demonstrates the potential of implicit reconstruction for image-based 3D perception. Extensive ex periments conducted on the KITTI-3D benchmark and Waymo Open Dataset demonstrate the effectiveness of MonoNeRD. Codes are available at https://github.com/cskkxj k/MonoNeRD.
****
Monocular 3D Object Detection with Bounding Box Denoising in 3D by Perceiver
Xianpeng Liu, Ce Zheng, Kelvin B Cheng, Nan Xue, Guo-Jun Qi, Tianfu Wu; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp . 6436-6446
The main challenge of monocular 3D object detection is the accurate localization of 3D center. Motivated by a new and strong observation that this challenge can be remedied by a 3D-space local-grid search scheme in an ideal case, we propose a stage-wise approach, which combines the information flow from 2D-to-3D (3D bo unding box proposal generation with a single 2D image) and 3D-to-2D (proposal ve rification by denoising with 3D-to-2D contexts) in a top-down manner. Specifical ly, we first obtain initial proposals from off-the-shelf backbone monocular 3D d etectors. Then, we generate a 3D anchor space by local-grid sampling from the in itial proposals. Finally, we perform 3D bounding box denoising at the 3D-to-2D p roposal verification stage. To effectively learn discriminative features for den oising highly overlapped proposals, this paper presents a method of using the Pe rceiver I/O model to fuse the 3D-to-2D geometric information and the 2D appearan

ce information. With the encoded latent representation of a proposal, the verification head is implemented with a self-attention module. Our method, named as MonoXiver, is generic and can be easily adapted to any backbone monocular 3D detectors. Experimental results on the well-established KITTI dataset and the challenging large-scale Waymo dataset show that MonoXiver consistently achieves improvement with limited computation overhead.
****

Point-SLAM: Dense Neural Point Cloud-based SLAM
Erik Sandström, Yue Li, Luc Van Gool, Martin R. Oswald; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18433-18444
We propose a dense neural simultaneous localization and mapping (SLAM) approach for monocular RGBD input which anchors the features of a neural scene representation in a point cloud that is iteratively generated in an input-dependent data-driven manner. We demonstrate that both tracking and mapping can be performed with the same point-based neural scene representation by minimizing an RGBD-based re-rendering loss. In contrast to recent dense neural SLAM methods which anchor the scene features in a sparse grid, our point-based approach allows to dynamically adapt the anchor point density to the information density of the input. This strategy reduces runtime and memory usage in regions with fewer details and dedicates higher point density to resolve fine details. Our approach performs either better or competitive to existing dense neural RGBD SLAM methods in tracking, mapping and rendering accuracy on the Replica, TUM-RGBD and ScanNet datasets.
****

TrajectoryFormer: 3D Object Tracking Transformer with Predictive Trajectory Hypotheses
Xuesong Chen, Shaoshuai Shi, Chao Zhang, Benjin Zhu, Qiang Wang, Ka Chun Cheung, Simon See, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18527-18536
3D multi-object tracking (MOT) is vital for many applications including autonomous driving vehicles and service robots. With the commonly used tracking-by-detection paradigm, 3D MOT has made important progress in recent years. However, these methods only use the detection boxes of the current frame to obtain trajectory-box association results, which makes it impossible for the tracker to recover objects missed by the detector. In this paper, we present TrajectoryFormer, a novel point-cloud-based 3D MOT framework. To recover the missed object by detector, we generates multiple trajectory hypotheses with hybrid candidate boxes, including temporally predicted boxes and currentframe detection boxes, for trajectory-box association. The predicted boxes can propagate object's history trajectory information to the current frame and thus the network can tolerate short-term miss detection of the tracked objects. We combine long-term object motion feature and short-term object appearance feature to create per-hypothesis feature embedding, which reduces the computational overhead for spatial-temporal encoding. Additionally, we introduce a Global-Local Interaction Module to conduct information interaction among all hypotheses and models their spatial relations, leading to accurate estimation of hypotheses. Our TrajectoryFormer achieves state-of-the-art performance on the Waymo 3D MOT benchmarks.
****

Semantic-Aware Dynamic Parameter for Video Inpainting Transformer
Eunhye Lee, Jinsu Yoo, Yunjeong Yang, Sungyong Baik, Tae Hyun Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12949-12958
Recent learning-based video inpainting approaches have achieved considerable progress. However, they still cannot fully utilize semantic information within the video frames and predict improper scene layout, failing to restore clear object boundaries for mixed scenes. To mitigate this problem, we introduce a new transformer-based video inpainting technique that can exploit semantic information within the input and considerably improve reconstruction quality. In this study, we use the mixture-of-experts scheme and train multiple experts to handle mixed scenes, including various semantics. We leverage these multiple experts and produce locally (token-wise) different network parameters to achieve semantic-aware in

painting results. Extensive experiments on YouTube-VOS and DAVIS benchmark datasets demonstrate that, compared with existing conventional video inpainting approaches, the proposed method has superior performance in synthesizing visually pleasing videos with much clearer semantic structures and textures.
****
See More and Know More: Zero-shot Point Cloud Segmentation via Multi-modal Visual Data
Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, Yuexin Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21674-21684
Zero-shot point cloud segmentation aims to make deep models capable of recognizing novel objects in point cloud that are unseen in the training phase. Recent trends favor the pipeline which transfers knowledge from seen classes with labels to unseen classes without labels. They typically align visual features with semantic features obtained from word embedding by the supervision of seen classes' annotations. However, point cloud contains limited information to fully match with semantic features. In fact, the rich appearance information of images is a natural complement to the textureless point cloud, which is not well explored in previous literature. Motivated by this, we propose a novel multi-modal zero-shot learning method to better utilize the complementary information of point clouds and images for more accurate visual-semantic alignment. Extensive experiments are performed in two popular benchmarks, i.e, SemanticKITTI and nuScenes, and our method outperforms current SOTA methods with 52% and 49% improvement on average for unseen class mIoU, respectively.
****
SKED: Sketch-guided Text-based 3D Editing
Aryan Mikaeili, Or Perel, Mehdi Safaee, Daniel Cohen-Or, Ali Mahdavi-Amiri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14607-14619
Text-to-image diffusion models are gradually introduced into computer graphics, recently enabling the development of Text-to-3D pipelines in an open domain. However, for interactive editing purposes, local manipulations of content through a simplistic textual interface can be arduous. Incorporating user guided sketches with Text-to-image pipelines offers users more intuitive control. Still, as state-of-the-art Text-to-3D pipelines rely on optimizing Neural Radiance Fields (NeRF) through gradients from arbitrary rendering views, conditioning on sketches is not straightforward. In this paper, we present SKED, a technique for editing 3D shapes represented by NeRFs. Our technique utilizes as few as two guiding sketches from different views to alter an existing neural field. The edited region respects the prompt semantics through a pre-trained diffusion model. To ensure the generated output adheres to the provided sketches, we propose novel loss functions to generate the desired edits while preserving the density and radiance of the base instance. We demonstrate the effectiveness of our proposed method through several qualitative and quantitative experiments. https://sked-paper.github.io/
****
WaveIPT: Joint Attention and Flow Alignment in the Wavelet domain for Pose Transfer
Liyuan Ma, Tingwei Gao, Haitian Jiang, Haibin Shen, Kejie Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7215-7225
Human pose transfer aims to generate a new image of the source person in a target pose. Among the existing methods, attention and flow are two of the most popular and effective approaches. Attention excels in preserving the semantic structure of the source image, which is more reflected in the low-frequency domain. Contrastively, flow is better at retaining fine-grained texture details in the high-frequency domain. To leverage the advantages of both attention and flow simultaneously, we propose Wavelet-aware Image-based Pose Transfer (WaveIPT) to fuse the attention and flow in the wavelet domain. To improve the fusion effect and avoid interference from irrelevant information between different frequencies, WaveI

PT first applies Intra-scale Local Correlation (ILC) to adaptively fuse attention and flow in the same scale according to their strengths in low and high-frequency domains, and then uses Inter-scale Feature Interaction (IFI) module to explore inter-scale frequency features for effective information transfer across different scales. We further introduce an effective Progressive Flow Regularization to alleviate the challenges of flow estimation under large pose differences. Our experiments on the DeepFashion dataset demonstrate that WaveIPT achieves a new state-of-the-art in terms of FID and LPIPS, with improvements of 4.97% and 3.89%, respectively.

****

Editable Image Geometric Abstraction via Neural Primitive Assembly

Ye Chen, Bingbing Ni, Xuanhong Chen, Zhangli Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23514-23523

This work explores a novel image geometric abstraction paradigm based on assembly out of a pool of pre-defined simple parametric primitives (i.e., triangle, rectangle, circle and semicircle), facilitating controllable shape editing in images. While cast as a mixed combinatorial and continuous optimization problem, the above task is approximately reformulated within a token translation neural framework that simultaneously outputs primitive assignments and corresponding transformation and color parameters in an image-to-set manner, thus bypassing complex/non-differentiable graph-matching iterations. To relax the searching space and address the gradient vanishing issue, a novel Neural Soft Assignment scheme that well explores the quasi-equivalence between the assignment in Bipartite b-Matching and opacity-aware weighted multiple rasterization combination is introduced, drastically reducing the optimization complexity. Without ground-truth image abstraction labeling (i.e., vectorized representation), the whole pipeline is end-to-end trainable in a self-supervised manner, based on the linkage of differentiable rasterization techniques. Extensive experiments on several datasets well demonstrate that our framework is able to predict highly compelling vectorized geometric abstraction results with a combination of ONLY four simple primitives, also with VERY straightforward shape editing capability by simple replacement of primitive type, compared to previous image abstraction and image vectorization methods.

****

Homeomorphism Alignment for Unsupervised Domain Adaptation

Lihua Zhou, Mao Ye, Xiatian Zhu, Siying Xiao, Xu-Qian Fan, Ferrante Neri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18699-18710

Existing unsupervised domain adaptation (UDA) methods rely on aligning the features from the source and target domains explicitly or implicitly in a common space (i.e., the domain invariant space). Explicit distribution matching ignores the discriminability of learned features, while the implicit counterpart such as self-supervised learning suffers from pseudo-label noises. With distribution alignment, it is challenging to acquire a common space which maintains fully the discriminative structure of both domains. In this work, we propose a novel HomeomorphisM Alignment (HMA) approach characterized by aligning the source and target data in two separate spaces. Specifically, an invertible neural network based homeomorphism is constructed. Distribution matching is then used as a sewing up tool for connecting this homeomorphism mapping between the source and target feature spaces. Theoretically, we show that this mapping can preserve the data topological structure (e.g., the cluster/group structure). This property allows for more discriminative model adaptation by leveraging both the original and transformed features of source data in a supervised manner, and those of target domain in an unsupervised manner (e.g., prediction consistency). Extensive experiments demonstrate that our method can achieve the state-of-the-art results. Code is released at https://github.com/buerzlh/HMA.

****

MBPTrack: Improving 3D Point Cloud Tracking with Memory Networks and Box Priors

Tian-Xing Xu, Yuan-Chen Guo, Yu-Kun Lai, Song-Hai Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9911-9920

3D single object tracking has been a crucial problem for decades with numerous applications such as autonomous driving. Despite its wide-ranging use, this task remains challenging due to the significant appearance variation caused by occlusion and size differences among tracked targets. To address these issues, we present MBPTrack, which adopts a Memory mechanism to utilize past information and formulates localization in a coarse-to-fine scheme using Box Priors given in the first frame. Specifically, past frames with targetness masks serve as an external memory, and a transformer-based module propagates tracked target cues from the memory to the current frame. To precisely localize objects of all sizes, MBPTrack first predicts the target center via Hough voting. By leveraging box priors given in the first frame, we adaptively sample reference points around the target center that roughly cover the target of different sizes. Then, we obtain dense feature maps by aggregating point features into the reference points, where localization can be performed more effectively. Extensive experiments demonstrate that MBPTrack achieves state-of-the-art performance on KITTI, nuScenes and Waymo Open Dataset, while running at 50 FPS on a single RTX3090 GPU.

****

Novel-View Synthesis and Pose Estimation for Hand-Object Interaction from Sparse Views

Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, Hongan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15100-15111

Hand-object interaction understanding and the barely addressed novel view synthesis are highly desired in the immersive communication, whereas it is challenging due to the high deformation of hand and heavy occlusions between hand and object. In this paper, we propose a neural rendering and pose estimation system for hand-object interaction from sparse views, which can also enable 3D hand-object interaction editing. We share the inspiration from recent scene understanding work that shows a scene specific model built beforehand can significantly improve and unblock vision tasks especially when inputs are sparse, and extend it to the dynamic hand-object interaction scenario and propose to solve the problem in two stages. We first learn the shape and appearance prior knowledge of hands and objects separately with the neural representation at the offline stage. During the online stage, we design a rendering-based joint model fitting framework to understand the dynamic hand-object interaction with the pre-built hand and object models as well as interaction priors, which thereby overcomes penetration and separation issues between hand and object and also enables novel view synthesis. In order to get stable contact during the hand-object interaction process in a sequence, we propose a stable contact loss to make the contact region to be consistent. Experiments demonstrate that our method outperforms the state-of-the-art methods. Code and dataset are available in project webpage https://iscas3dv.github.io/HO-NeRF.

****

EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes

Jingyuan Yang, Qirui Huang, Tingting Ding, Dani Lischinski, Danny Cohen-Or, Hui Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20383-20394

Visual Emotion Analysis (VEA) aims at predicting people's emotional responses to visual stimuli. This is a promising, yet challenging, task in affective computing, which has drawn increasing attention in recent years. Most of the existing work in this area focuses on feature design, while little attention has been paid to dataset construction. In this work, we introduce EmoSet, the first large-scale visual emotion dataset annotated with rich attributes, which is superior to existing datasets in four aspects: scale, annotation richness, diversity, and data balance. EmoSet comprises 3.3 million images in total, with 118,102 of these images carefully labeled by human annotators, making it five times larger than the largest existing dataset. EmoSet includes images from social networks, as well as artistic images, and it is well balanced between different emotion categories. Motivated by psychological studies, in addition to emotion category, each image is also annotated with a set of describable emotion attributes: brightness, c

olorfulness, scene type, object class, facial expression, and human action, which can help understand visual emotions in a precise and interpretable way. The relevance of these emotion attributes is validated by analyzing the correlations between them and visual emotion, as well as by designing an attribute module to help visual emotion recognition. We believe EmoSet will bring some key insights and encourage further research in visual emotion analysis and understanding. Project page: https://vcc.tech/EmoSet.
****
Distilling from Similar Tasks for Transfer Learning on a Budget
Kenneth Borup, Cheng Perng Phoo, Bharath Hariharan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11431-11441
We address the challenge of getting efficient yet accurate recognition systems with limited labels. While recognition models improve with model size and amount of data, many specialized applications of computer vision have severe resource constraints both during training and inference. Transfer learning is an effective solution for training with few labels, however often at the expense of a computationally costly fine-tuning of large base models. We propose to mitigate this unpleasant trade-off between compute and accuracy via semi-supervised cross-domain distillation from a set of diverse source models. Initially, we show how to use task similarity metrics to select a single suitable source model to distill from, and that a good selection process is imperative for good downstream performance of a target model. We dub this approach DistillNearest. Though effective, DistillNearest assumes a single source model matches the target task, which is not always the case. To alleviate this, we propose a weighted multi-source distillation method to distill multiple source models trained on different domains weighted by their relevance for the target task into a single efficient model (named DistillWeighted). Our methods need no access to source data and merely need features and pseudo-labels of the source models. When the goal is accurate recognition under computational constraints, both DistillNearest and DistillWeighted approaches outperform both transfer learning from strong ImageNet initializations as well as state-of-the-art semi-supervised techniques such as FixMatch. Averaged over 8 diverse target tasks our multi-source method outperforms the baselines by 5.6%-points and 4.5%-points, respectively.
****
Self-Supervised Burst Super-Resolution
Goutam Bhat, Michaël Gharbi, Jiawen Chen, Luc Van Gool, Zhihao Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10605-10614
We introduce a self-supervised training strategy for burst super-resolution that only uses noisy low-resolution bursts during training. Our approach eliminates the need to carefully tune synthetic data simulation pipelines, which often do not match real-world image statistics. Compared to weakly-paired training strategies, which require noisy smartphone burst photos of static scenes, paired with a clean reference obtained from a tripod-mounted DSLR camera, our approach is more scalable, and avoids the color mismatch between the smartphone and DSLR. To achieve this, we propose a new self-supervised objective that uses a forward imaging model to recover a high-resolution image from aliased high frequencies in the burst. Our approach does not require any manual tuning of the forward model's parameters; we learn them from data. Furthermore, we show our training strategy is robust to dynamic scene motion in the burst, which enables training burst super-resolution models using in-the-wild data. Extensive experiments on real and synthetic data show that, despite only using noisy bursts during training, models trained with our self-supervised strategy match, and sometimes surpass, the quality of fully-supervised baselines trained with synthetic data or weakly-paired ground-truth. Finally, we show our training strategy is general using four different burst super-resolution architectures.
****
Class-relation Knowledge Distillation for Novel Class Discovery
Peiyan Gu, Chuyu Zhang, Ruijie Xu, Xuming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16474-16483

We tackle the problem of novel class discovery, which aims to learn novel classe
s without supervision based on labeled data from known classes. A key challenge
lies in transferring the knowledge in the known-class data to the learning of no
vel classes. Previous methods mainly focus on building a shared representation s
pace for knowledge transfer and often ignore modeling class relations. To addres
s this, we introduce a class relation representation for the novel classes based
 on the predicted class distribution of a model trained on known classes. Empiri
cally, we find that such class relation becomes less informative during typical
discovery training. To prevent such information loss, we propose a novel knowled
ge distillation framework, which utilizes our class-relation representation to r
egularize the learning of novel classes. In addition, to enable a flexible knowl
edge distillation scheme for each data point in novel classes, we develop a lear
nable weighting function for the regularization, which adaptively promotes knowl
edge transfer based on the semantic similarity between the novel and known class
es. To validate the effectiveness and generalization of our method, we conduct e
xtensive experiments on multiple benchmarks, including CIFAR100, Stanford Cars,
CUB, and FGVC-Aircraft datasets. Our results demonstrate that the proposed metho
d outperforms the previous state-of-the-art methods by a significant margin on a
lmost all benchmarks.
****
PARTNER: Level up the Polar Representation for LiDAR 3D Object Detection
Ming Nie, Yujing Xue, Chunwei Wang, Chaoqiang Ye, Hang Xu, Xinge Zhu, Qingqiu Hu
ang, Michael Bi Mi, Xinchao Wang, Li Zhang; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 3801-3813
Recently, polar-based representation has shown promising properties in perceptua
l tasks. In addition to Cartesian-based approaches, which separate point clouds
unevenly, representing point clouds as polar grids has been recognized as an alt
ernative due to (1) its advantage in robust performance under different resoluti
ons and (2) its superiority in streaming-based approaches. However, state-of-the
-art polar-based detection methods inevitably suffer from the feature distortion
 problem because of the non-uniform division of polar representation, resulting
in a non-negligible performance gap compared to Cartesian-based approaches. To t
ackle this issue, we present PARTNER, a novel 3D object detector in the polar co
ordinate. PARTNER alleviates the dilemma of feature distortion with global repre
sentation re-alignment and facilitates the regression by introducing instance-le
vel geometric information into the detection head. Extensive experiments show ov
erwhelming advantages in streaming-based detection and different resolutions. Fu
rthermore, our method outperforms the previous polar-based works with remarkable
 margins of 3.68% and 9.15% on Waymo and ONCE validation set, thus achieving com
petitive results over the state-of-the-art methods.
****
Data-Free Class-Incremental Hand Gesture Recognition
Shubhra Aich, Jesus Ruiz-Santaquiteria, Zhenyu Lu, Prachi Garg, K J Joseph, Alva
ro Fernandez Garcia, Vineeth N Balasubramanian, Kenrick Kin, Chengde Wan, Necati
 Cihan Camgoz, Shugao Ma, Fernando De la Torre; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2023, pp. 20958-20967
This paper investigates data-free class-incremental learning (DFCIL)
 for hand gesture recognition from 3D skeleton sequences.
 In this class-incremental learning (CIL) setting, while incrementally
 registering the new classes, we do not have access to the training
 samples (i.e. data-free) of the already known classes due to privacy.
 Existing DFCIL methods primarily focus on various forms of
 knowledge distillation for model inversion to mitigate
 catastrophic forgetting. Unlike SOTA methods,
 we delve deeper into the choice of the best samples for inversion.
 Inspired by the well-grounded theory of max-margin classification,
 we find that the best samples tend to lie close to the approximate
 decision boundary within a reasonable margin. To this end,
 we propose BOAT-MI -- a simple and effective boundary-aware prototypical
 sampling mechanism for model inversion for DFCIL.

Our sampling scheme outperforms SOTA methods significantly
on two 3D skeleton gesture datasets, the publicly available
SHREC 2017, and EgoGesture3D -- which we extract from a publicly
available RGBD dataset. Both our codebase and the EgoGesture3D
skeleton dataset are publicly available: https://github.com/humansensinglab/dfcil-hgr
****
Corrupting Neuron Explanations of Deep Visual Features
Divyansh Srivastava, Tuomas Oikarinen, Tsui-Wei Weng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1877-1886
The inability of DNNs to explain their black-box behavior has led to a recent surge of explainability methods. However, there are growing concerns that these explainability methods are not robust and trustworthy. In this work, we perform the first robustness analysis of Neuron Explanation Methods under a unified pipeline and show that these explanations can be significantly corrupted by random noises and well-designed perturbations added to their probing data. We find that even adding small random noise with a standard deviation of 0.02 can already change the assigned concepts of up to 28% neurons in the deeper layers. Furthermore, we devise a novel corruption algorithm and show that our algorithm can manipulate the explanation of more than 80% neurons by poisoning less than 10% of probing data. This raises the concern of trusting Neuron Explanation Methods in real-life safety and fairness critical applications.
****
PNI : Industrial Anomaly Detection using Position and Neighborhood Information
Jaehyeok Bae, Jae-Han Lee, Seyun Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6373-6383
Because anomalous samples cannot be used for training,
many anomaly detection and localization methods use
pre-trained networks and non-parametric modeling to estimate
encoded feature distribution. However, these methods
neglect the impact of position and neighborhood information
on the distribution of normal features. To overcome
this, we propose a new algorithm, PNI, which estimates
the normal distribution using conditional probability given
neighborhood features, modeled with a multi-layer perceptron
network. Moreover, position information is utilized by
creating a histogram of representative features at each position.
Instead of simply resizing the anomaly map, the proposed
method employs an additional refine network trained
on synthetic anomaly images to better interpolate and account
for the shape and edge of the input image. We conducted
experiments on the MVTec AD benchmark dataset
and achieved state-of-the-art performance, with 99.56%
and 98.98% AUROC scores in anomaly detection and localization,
respectively. Code is available at https://github.com/wogur110/PNI_Anomaly_Detection.
****
PC-Adapter: Topology-Aware Adapter for Efficient Domain Adaption on Point Clouds with Rectified Pseudo-label
Joonhyung Park, Hyunjin Seo, Eunho Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11530-11540
Understanding point clouds captured from the real-world is challenging due to shifts in data distribution caused by varying object scales, sensor angles, and self-occlusion. Prior works have addressed this issue by combining recent learning principles such as self-supervised learning, self-training and adversarial training, which leads to to significant computational overhead. Toward succinct yet powerful domain adaptation for point clouds, we revisit the unique challenges of point cloud data under domain shift scenarios and discover the importance of the global geometry of source data and trends of target pseudo-labels biased to the source label distribution. Motivated by our observations, we propose an adapte

r-guided domain adaptation method, PC-Adapter, that preserves the global shape information of the source domain using an attention-based adapter, while learning the local characteristics of the target domain via another adapter equipped with graph convolution. Additionally, we propose a novel pseudo-labeling strategy resilient to the classifier bias by adjusting confidence scores using their class-wise confidence distributions to consider relative confidences. Our method demonstrates superiority over baselines on various domain shift settings in benchmark datasets - PointDA, GraspNetPC, and PointSegDA.
****
Cyclic Test-Time Adaptation on Monocular Video for 3D Human Mesh Reconstruction
Hyeongjin Nam, Daniel Sungho Jung, Yeonguk Oh, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14829-14839
Despite recent advances in 3D human mesh reconstruction, domain gap between training and test data is still a major challenge. Several prior works tackle the domain gap problem via test-time adaptation that fine-tunes a network relying on 2D evidence (e.g., 2D human keypoints) from test images. However, the high reliance on 2D evidence during adaptation causes two major issues. First, 2D evidence induces depth ambiguity, preventing the learning of accurate 3D human geometry. Second, 2D evidence is noisy or partially non-existent during test time, and such imperfect 2D evidence leads to erroneous adaptation. To overcome the above issues, we introduce CycleAdapt, which cyclically adapts two networks: a human mesh reconstruction network (HMRNet) and a human motion denoising network (MDNet), given a test video. In our framework, to alleviate high reliance on 2D evidence, we fully supervise HMRNet with generated 3D supervision targets by MDNet. Our cyclic adaptation scheme progressively elaborates the 3D supervision targets, which compensate for imperfect 2D evidence. As a result, our CycleAdapt achieves state-of-the-art performance compared to previous test-time adaptation methods. The codes are available in here: https://github.com/hygenie1228/CycleAdapt_RELEASE.
****
2D3D-MATR: 2D-3D Matching Transformer for Detection-Free Registration Between Images and Point Clouds
Minhao Li, Zheng Qin, Zhirui Gao, Renjiao Yi, Chenyang Zhu, Yulan Guo, Kai Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14128-14138
The commonly adopted detect-then-match approach to registration finds difficulties in the cross-modality cases due to the incompatible keypoint detection and inconsistent feature description. We propose, 2D3D-MATR, a detection-free method for accurate and robust registration between images and point clouds. Our method adopts a coarse-to-fine pipeline where it first computes coarse correspondences between downsampled patches of the input image and the point cloud and then extends them to form dense correspondences between pixels and points within the patch region. The coarse-level patch matching is based on transformer which jointly learns global contextual constraints with self-attention and cross-modality correlations with cross-attention. To resolve the scale ambiguity in patch matching, we construct a multi-scale pyramid for each image patch and learn to find for each point patch the best matching image patch at a proper resolution level. Extensive experiments on two public benchmarks demonstrate that 2D3D-MATR outperforms the previous state-of-the-art P2-Net by around 20 percentage points on inlier ratio and over 10 points on registration recall. Our code and models will be publicly released.
****
Mixed Neural Voxels for Fast Multi-view Video Synthesis
Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, Huaping Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19706-19716
Synthesizing high-fidelity videos from real-world multiview input is challenging due to the complexities of real-world environments and high-dynamic movements. Previous works based on neural radiance fields have demonstrated high-quality reconstructions of dynamic scenes. However, training such models on real-world sce

nes is time-consuming, usually taking days or weeks. In this paper, we present a novel method named MixVoxels to efficiently represent dynamic scenes which leads to fast training and rendering speed. The proposed MixVoxels represents the 4D dynamic scenes as a mixture of static and dynamic voxels and processes them with different networks. In this way, the computation of the required modalities for static voxels can be processed by a lightweight model, which essentially reduces
the amount of computation as many daily dynamic scenes are dominated by static backgrounds. To distinguish the two kinds of voxels, we propose a novel variation field to estimate the temporal variance of each voxel. For the dynamic representations, we design an inner-product time query method to efficiently query multiple time steps, which is essential to recover the high-dynamic movements. As a result, with 15 minutes of training for dynamic scenes with inputs of 300-frame videos, MixVoxels achieves better PSNR than previous methods. For rendering, MixVoxels can render a novel view video with 1K resolution at 37 fps. Codes and trained models are available at https://github.com/fengres/mixvoxels.
****
Bidirectionally Deformable Motion Modulation For Video-based Human Pose Transfer
Wing-Yin Yu, Lai-Man Po, Ray C.C. Cheung, Yuzhi Zhao, Yu Xue, Kun Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7502-7512
Video-based human pose transfer is a video-to-video generation task that animates a plain source human image based on a series of target human poses. Considering the difficulties in transferring highly structural patterns on the garments and discontinuous poses, existing methods often generate unsatisfactory results such as distorted textures and flickering artifacts. To address these issues, we propose a novel Deformable Motion Modulation (DMM) that utilizes geometric kernel offset with adaptive weight modulation to simultaneously perform feature alignment and style transfer. Different from normal style modulation used in style transfer, the proposed modulation mechanism adaptively reconstructs smoothed frames from style codes according to the object shape through an irregular receptive field of view. To enhance the spatio-temporal consistency, we leverage bidirectional propagation to extract the hidden motion information from a warped image sequence generated by noisy poses. The proposed feature propagation significantly enhances the motion prediction ability by forward and backward propagation. Both quantitative and qualitative experimental results demonstrate superiority over the state-of-the-arts in terms of image fidelity and visual continuity. The source code is publicly available at github.com/rocketappslab/bdmm.
****
Harvard Glaucoma Detection and Progression: A Multimodal Multitask Dataset and Generalization-Reinforced Semi-Supervised Learning
Yan Luo, Min Shi, Yu Tian, Tobias Elze, Mengyu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20471-20482
Glaucoma is the number one cause of irreversible blindness globally. A major challenge for accurate glaucoma detection and progression forecasting is the bottleneck of limited labeled patients with the state-of-the-art (SOTA) 3D retinal imaging data of optical coherence tomography (OCT). To address the data scarcity issue, this paper proposes two solutions. First, we develop a novel generalization-reinforced semi-supervised learning (SSL) model called pseudo supervisor to optimally utilize unlabeled data. Compared with SOTA models, the proposed pseudo supervisor optimizes the policy of predicting pseudo labels with unlabeled samples to improve empirical generalization. Our pseudo supervisor model is evaluated with two clinical tasks consisting of glaucoma detection and progression forecasting. The progression forecasting task is evaluated both unimodally and multimodally. Our pseudo supervisor model demonstrates superior performance than SOTA SSL comparison models. Moreover, our model also achieves the best results on the publicly available LAG fun- dus dataset. Second, we introduce the Harvard Glaucoma Detection and Progression (Harvard-GDP) Dataset, a multimodal multitask dataset that includes data from 1,000 patients with OCT imaging data, as well as labels for glaucoma detection and progression. This is the largest glaucoma detection

dataset with 3D OCT imaging data and the first glaucoma progression forecasting dataset that is publicly available. Detailed sex and racial analysis are pro- vided, which can be used by interested researchers for fairness learning studies. Our released dataset is benchmarked with several SOTA supervised CNN and transformer deep learning models. The dataset and code are made publicly available via https://ophai.hms.harvard.edu/ datasets/harvard-gdp1000.
****
Tracking Everything Everywhere All at Once
Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, Noah Snavely; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19795-19806
We present a new test-time optimization method for estimating dense and long-range motion from a video sequence. Prior optical flow or particle video tracking algorithms typically operate within limited temporal windows, struggling to track through occlusions and maintain global consistency of estimated motion trajectories. We propose a complete and globally consistent motion representation, dubbed OmniMotion, that allows for accurate, full-length motion estimation of every pixel in a video. OmniMotion represents a video using a quasi-3D canonical volume and performs pixel-wise tracking via bijections between local and canonical space. This representation allows us to ensure global consistency, track through occlusions, and model any combination of camera and object motion. Extensive evaluations on the TAP-Vid benchmark and real-world footage show that our approach outperforms prior state-of-the-art methods by a large margin both quantitatively and qualitatively.
****
Group Pose: A Simple Baseline for End-to-End Multi-Person Pose Estimation
Huan Liu, Qiang Chen, Zichang Tan, Jiang-Jiang Liu, Jian Wang, Xiangbo Su, Xiaolong Li, Kun Yao, Junyu Han, Errui Ding, Yao Zhao, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15029-15038
In this paper, we study the problem of end-to-end multi-person pose estimation. State-of-the-art solutions adopt the DETR-like framework, and mainly develop the complex decoder, e.g., regarding pose estimation as keypoint box detection and combining with human detection in ED-Pose, hierarchically predicting with pose decoder and joint (keypoint) decoder in PETR.

We present a simple yet effective transformer approach, named Group Pose. We simply regard K-keypoint pose estimation as predicting a set of NxK keypoint positions, each from a keypoint query, as well as representing each pose with an instance query for scoring N pose predictions.

Motivated by the intuition that the interaction, among across-instance queries of different types, is not directly helpful, we make a simple modification to decoder self-attention. We replace single self-attention over all the Nx(K+1) queries with two subsequent group self-attentions: (i) N within-instance self-attention, with each over K keypoint queries and one instance query, and (ii) (K+1) same-type across-instance self-attention, each over N queries of the same type. The resulting decoder removes the interaction among across-instance type-different queries, easing the optimization and thus improving the performance. Experimental results on MS COCO and CrowdPose show that our approach without human box supervision is superior to previous methods with complex decoders, and even is slightly better than ED-Pose that uses human box supervision. Code is available.
****
Objects Do Not Disappear: Video Object Detection by Single-Frame Object Location Anticipation
Xin Liu, Fatemeh Karimi Nejadasl, Jan C. van Gemert, Olaf Booij, Silvia L. Pintea; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6950-6961
Objects in videos are typically characterized by continuous smooth motion. We exploit continuous smooth motion in three ways. 1) Improved accuracy by using obje

ct motion as an additional source of supervision, which we obtain by anticipating object locations from a static keyframe. 2) Improved efficiency by only doing the expensive feature computations on a small subset of all frames. Because neighboring video frames are often redundant, we only compute features for a single static keyframe and predict object locations in subsequent frames. 3) Reduced annotation cost, where we only annotate the keyframe and use smooth pseudo-motion between keyframes. We demonstrate computational efficiency, annotation efficiency, and improved mean average precision compared to the state-of-the-art on four datasets: ImageNet VID, EPIC KITCHENS-55, YouTube-BoundingBoxes and Waymo Open dataset. Our source code is available at https://github.com/L-KID/Videoobject-detection-by-location-anticipation.
****
CauSSL: Causality-inspired Semi-supervised Learning for Medical Image Segmentation
Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, Pheng-Ann Heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21426-21437
Semi-supervised learning (SSL) has recently demonstrated great success in medical image segmentation, significantly enhancing data efficiency with limited annotations. However, despite its empirical benefits, there are still concerns in the literature about the theoretical foundation and explanation of semi-supervised segmentation. To explore this problem, this study first proposes a novel causal diagram to provide a theoretical foundation for the mainstream semi-supervised segmentation methods. Our causal diagram takes two additional intermediate variables into account, which are neglected in previous work. Drawing from this proposed causal diagram, we then introduce a causality-inspired SSL approach on top of co-training frameworks called CauSSL, to improve SSL for medical image segmentation. Specifically, we first point out the importance of algorithmic independence between two networks or branches in SSL, which is often overlooked in the literature. We then propose a novel statistical quantification of the uncomputable algorithmic independence and further enhance the independence via a min-max optimization process. Our method can be flexibly incorporated into different existing SSL methods to improve their performance. Our method has been evaluated on three challenging medical image segmentation tasks using both 2D and 3D network architectures and has shown consistent improvements over state-of-the-art methods. Our code is publicly available at: https://github.com/JuzhengMiao/CauSSL.
****
ChartReader: A Unified Framework for Chart Derendering and Comprehension without Heuristic Rules
Zhi-Qi Cheng, Qi Dai, Alexander G. Hauptmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22202-22213
Charts are a powerful tool for visually conveying complex data, but their comprehension poses a challenge due to the diverse chart types and intricate components. Existing chart comprehension methods suffer from either heuristic rules or an over-reliance on OCR systems, resulting in suboptimal performance. To address these issues, we present ChartReader, a unified framework that seamlessly integrates chart derendering and comprehension tasks. Our approach includes a transformer-based chart component detection module and an extended pre-trained vision-language model for chart-to-X tasks. By learning the rules of charts automatically from annotated datasets, our approach eliminates the need for manual rule-making, reducing effort and enhancing accuracy. We also introduce a data variable replacement technique and extend the input and position embeddings of the pre-trained model for cross-task training. We evaluate ChartReader on Chart-to-Table, ChartQA, and Chart-to-Text tasks, demonstrating its superiority over existing methods. Our proposed framework can significantly reduce the manual effort involved in chart analysis, providing a step towards a universal chart understanding model. Moreover, our approach offers opportunities for plug-and-play integration with mainstream LLMs such as T5 and TaPas, extending their capability to chart comprehension tasks.
****

Learning from Semantic Alignment between Unpaired Multiviews for Egocentric Vide o Recognition
Qitong Wang, Long Zhao, Liangzhe Yuan, Ting Liu, Xi Peng; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3307-3317
We are concerned with a challenging scenario in unpaired multiview video learnin g. In this case, the model aims to learn comprehensive multiview representations while the cross-view semantic information exhibits variations. We propose Seman tics-based Unpaired Multiview Learning (SUM-L) to tackle this unpaired multiview learning problem. The key idea is to build cross-view pseudo-pairs and do view-invariant alignment by leveraging the semantic information of videos. To facilit ate the data efficiency of multiview learning, we further perform video-text ali gnment for first-person and third-person videos, to fully leverage the semantic knowledge to improve video representations. Extensive experiments on multiple be nchmark datasets verify the effectiveness of our framework. Our method also outp erforms multiple existing view-alignment methods, under the more challenging sce nario than typical paired or unpaired multimodal or multiview learning. Our code is available at https://github.com/wqtwjt1996/SUM-L.
****
Neural LiDAR Fields for Novel View Synthesis
Shengyu Huang, Zan Gojcic, Zian Wang, Francis Williams, Yoni Kasten, Sanja Fidle r, Konrad Schindler, Or Litany; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2023, pp. 18236-18246
We present Neural Fields for LiDAR (NFL), a method to optimise a neural field sc ene representation from LiDAR measurements, with the goal of synthesizing realis tic LiDAR scans from novel viewpoints. NFL combines the rendering power of neura l fields with a detailed, physically motivated model of the LiDAR sensing proces s, thus enabling it to accurately reproduce key sensor behaviors like beam diver gence, secondary returns, and ray dropping. We evaluate NFL on synthetic and rea l LiDAR scans and show that it outperforms explicit reconstruct-then-simulate me thods as well as other NeRF-style methods on LiDAR novel view synthesis task. Mo reover, we show that the improved realism of the synthesized views narrows the d omain gap to real scans and translates to better registration and semantic segme ntation performance.
****
Source-free Depth for Object Pop-out
Zongwei WU, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric D emonceaux, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1032-1042
Depth cues are known to be useful for visual perception. However, direct measure ment of depth is often impracticable. Fortunately, though, modern learning-based methods offer promising depth maps by inference in the wild. In this work, we a dapt such depth inference models for object segmentation using the objects' "pop -out" prior in 3D. The "pop-out" is a simple composition prior that assumes obje cts reside on the background surface. Such compositional prior allows us to reas on about objects in the 3D space. More specifically, we adapt the inferred depth maps such that objects can be localized using only 3D information. Such separat ion, however, requires knowledge about contact surface which we learn using the weak supervision of the segmentation mask. Our intermediate representation of co ntact surface, and thereby reasoning about objects purely in 3D, allows us to be tter transfer the depth knowledge into semantics. The proposed adaptation method uses only the depth model without needing the source data used for training, ma king the learning process efficient and practical. Our experiments on eight data sets of two challenging tasks, namely salient object detection and camouflaged o bject detection, consistently demonstrate the benefit of our method in terms of both performance and generalizability. The source code is publicly available at https://github.com/Zongwei97/PopNet.
****
Token-Label Alignment for Vision Transformers
Han Xiao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/ CVF International Conference on Computer Vision (ICCV), 2023, pp. 5495-5504

Data mixing strategies (e.g., CutMix) have shown the ability to greatly improve the performance of convolutional neural networks (CNNs). They mix two images as inputs for training and assign them with a mixed label with the same ratio. While they are shown effective for vision transformers (ViTs), we identify a token fluctuation phenomenon that has suppressed the potential of data mixing strategies. We empirically observe that the contributions of input tokens fluctuate as forward propagating, which might induce a different mixing ratio in the output tokens. The training target computed by the original data mixing strategy can thus be inaccurate, resulting in less effective training. To address this, we propose a token-label alignment (TL-Align) method to trace the correspondence between transformed tokens and the original tokens to maintain a label for each token. We reuse the computed attention at each layer for efficient token-label alignment, introducing only negligible additional training costs. Extensive experiments demonstrate that our method improves the performance of ViTs on image classification, semantic segmentation, objective detection, and transfer learning tasks.
****
Understanding 3D Object Interaction from a Single Image
Shengyi Qian, David F. Fouhey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21753-21763
Humans can easily understand a single image as depicting multiple potential objects permitting interaction. We use this skill to plan our interactions with the world and accelerate understanding new objects without engaging in interaction. In this paper, we would like to endow machines with the similar ability, so that intelligent agents can better explore the 3D scene or manipulate objects. Our approach is a transformer-based model that predicts the 3D location, physical properties and affordance of objects. To power this model, we collect a dataset with Internet videos, egocentric videos and indoor images to train and validate our approach. Our model yields strong performance on our data, and generalizes well to robotics data.
****
SkeleTR: Towards Skeleton-based Action Recognition in the Wild
Haodong Duan, Mingze Xu, Bing Shuai, Davide Modolo, Zhuowen Tu, Joseph Tighe, Alessandro Bergamo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13634-13644
We present SkeleTR, a new framework for skeleton-based action recognition. In contrast to prior work, which focuses mainly on controlled environments, we target in-the-wild scenarios that typically involve a variable number of people and various forms of interaction between people. SkeleTR works with a two-stage paradigm. It first models the intra-person skeleton dynamics for each skeleton sequence with graph convolutions, and then uses stacked Transformer encoders to capture person interactions that are important for action recognition in the wild. To mitigate the negative impact of inaccurate skeleton associations, SkeleTR takes relative short skeleton sequences as input and increases the number of sequences. As a unified solution, SkeleTR can be directly applied to multiple skeleton-based action tasks, including video-level action classification, instance-level action detection, and group-level activity recognition. It also enables transfer learning and joint training across different action tasks and datasets, which results in performance improvement. When evaluated on various skeleton-based action recognition benchmarks, SkeleTR achieves the state-of-the-art performance.
****
Learning Gabor Texture Features for Fine-Grained Recognition
Lanyun Zhu, Tianrun Chen, Jianxiong Yin, Simon See, Jun Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1621-1631
Extracting and using class-discriminative features is critical for fine-grained recognition. Existing works have demonstrated the possibility of applying deep CNNs to exploit features that distinguish similar classes. However, CNNs suffer from problems including frequency bias and loss of detailed local information, which restricts the performance of recognizing fine-grained categories. To address the challenge, we propose a novel texture branch as complimentary to the CNN branch for feature extraction. We innovatively utilize Gabor filters as a powerful

extractor to exploit texture features, motivated by the capability of Gabor filters in effectively capturing multi-frequency features and detailed local information. We implement several designs to enhance the effectiveness of Gabor filters, including imposing constraints on parameter values and developing a learning method to determine the optimal parameters. Moreover, we introduce a statistical feature extractor to utilize informative statistical information from the signals captured by Gabor filters, and a gate selection mechanism to enable efficient computation by only considering qualified regions as input for texture extraction. Through the integration of features from the Gabor-filter-based texture branch and CNN-based semantic branch, we achieve comprehensive information extraction. We demonstrate the efficacy of our method on multiple datasets, including CUB-200-2011, NA-bird, Stanford Dogs, and GTOS-mobile. State-of-the-art performance is achieved using our approach.
****
Weakly-Supervised Action Localization by Hierarchically-Structured Latent Attention Modeling
Guiqin Wang, Peng Zhao, Cong Zhao, Shusen Yang, Jie Cheng, Luziwei Leng, Jianxing Liao, Qinghai Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10203-10213
Weakly-supervised action localization aims to recognize and localize action instancese in untrimmed videos with only video-level labels. Most existing models rely on multiple instance learning(MIL), where the predictions of unlabeled instances are supervised by classifying labeled bags. The MIL-based methods are relatively well studied with cogent performance achieved on classification but not on localization. Generally, they locate temporal regions by the video-level classification but overlook the temporal variations of feature semantics. To address this problem, we propose a novel attention-based hierarchically-structured latent model to learn the temporal variations of feature semantics. Specifically, our model entails two components, the first is an unsupervised change-points detection module that detects change-points by learning the latent representations of video features in a temporal hierarchy based on their rates of change, and the second is an attention-based classification model that selects the change-points of the foreground as the boundaries. To evaluate the effectiveness of our model, we conduct extensive experiments on two benchmark datasets, THUMOS-14 and ActivityNet-v1.3. The experiments show that our method outperforms current state-of-the-art methods, and even achieves comparable performance with fully-supervised methods.
****
Get3DHuman: Lifting StyleGAN-Human into a 3D Generative Model Using Pixel-Aligned Reconstruction Priors
Zhangyang Xiong, Di Kang, Derong Jin, Weikai Chen, Linchao Bao, Shuguang Cui, Xiaoguang Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9287-9297
Fast generation of high-quality 3D digital humans is important to a vast number of applications ranging from entertainment to professional concerns. Recent advances in differentiable rendering have enabled the training of 3D generative models without requiring 3D ground truths. However, the quality of the generated 3D humans still has much room to improve in terms of both fidelity and diversity. In this paper, we present Get3DHuman, a novel 3D human framework that can significantly boost the realism and diversity of the generated outcomes by only using a limited budget of 3D ground-truth data. Our key observation is that the 3D generator can profit from human-related priors learned through 2D human generators and 3D reconstructors. Specifically, we bridge the latent space of Get3DHuman with that of StyleGAN-Human via a specially-designed prior network, where the input latent code is mapped to the shape and texture feature volumes spanned by the pixel-aligned 3D reconstructor. The outcomes of the prior network are then leveraged as the supervisory signals for the main generator network. To ensure effective training, we further propose three tailored losses applied to the generated feature volumes and the intermediate feature maps. Extensive experiments demonstrate that Get3DHuman greatly outperforms the other state-of-the-art approaches an

d can support a wide range of applications including shape interpolation, shape re-texturing, and single-view reconstruction through latent inversion.
****
Query6DoF: Learning Sparse Queries as Implicit Shape Prior for Category-Level 6DoF Pose Estimation
Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, Wenyu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14055-14064
Category-level 6DoF object pose estimation intends to estimate the rotation, translation, and size of unseen objects. Many previous works use point clouds as a pre-learned shape prior to overcome intra-category variability. The shape prior is deformed to reconstruct instances' point clouds in canonical space and to build dense 3D-3D correspondences between the observed and reconstructed point clouds. However, the pre-learned shape prior is not jointly optimized with estimation networks, and they are trained with a surrogate objective. We propose a novel 6D pose estimation network, named Query6DoF, based on a series of category-specific sparse queries that represent the prior shape. Each query represents a shape component, and these queries are learnable embeddings that can be optimized together with the estimation network according to the point cloud reconstruction loss, the normalized object coordinate loss, and the 6d pose estimation loss. Query6DoF adopts a deformation-and-matching paradigm with attention, where the queries dynamically extract features from regions of interest using the attention mechanism and then directly regress results.
Furthermore, Query6DoF reduces computation overhead through the sparseness of the queries and the incorporation of a lightweight global information injection block. With the aforementioned design, Query6DoF achieves state-of-the-art (SOTA) pose estimation performance on the NOCS datasets. The source code and models are available at https://github.com/hustvl/Query6DoF.
****
Towards High-Quality Specular Highlight Removal by Leveraging Large-Scale Synthetic Data
Gang Fu, Qing Zhang, Lei Zhu, Chunxia Xiao, Ping Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12857-12865
This paper aims to remove specular highlights from a single object-level image. Although previous methods have made some progresses, their performance remains somewhat limited, particularly for real images with complex specular highlights. To this end, we propose a three-stage network to address them. Specifically, given an input image, we first decompose it into the albedo, shading, and specular residue components to estimate a coarse specular-free image. Then, we further refine the coarse result to alleviate its visual artifacts such as color distortion. Finally, we adjust the tone of the refined result to match the tone of the input as closely as possible. In addition, to facilitate network training and quantitative evaluation, we present a large-scale synthetic dataset of object-level images, covering diverse objects and illumination conditions. Extensive experiments illustrate that our network is able to generalize well to unseen real object-level images, and even produce good results for scene-level images with multiple background objects and complex lighting.
****
An Embarrassingly Simple Backdoor Attack on Self-supervised Learning
Changjiang Li, Ren Pang, Zhaohan Xi, Tianyu Du, Shouling Ji, Yuan Yao, Ting Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4367-4378
As a new paradigm in machine learning, self-supervised learning (SSL) is capable of learning high-quality representations of complex data without relying on labels. In addition to eliminating the need for labeled data, research has found that SSL improves the adversarial robustness over supervised learning since lacking labels makes it more challenging for adversaries to manipulate model predictions. However, the extent to which this robustness superiority generalizes to other types of attacks remains an open question. We explore this question in the context of backdoor attacks. Specifically, we design and evaluate CTRL, an embarras

singly simple yet highly effective self-supervised backdoor attack. By only polluting a tiny fraction of training data (<1%) with indistinguishable poisoning samples, CTRL causes any trigger-embedded input to be misclassified to the adversary's designated class with a high probability (>99%) at inference time. Our findings suggest that SSL and supervised learning are comparably vulnerable to backdoor attacks. More importantly, through the lens of CTRL, we study the inherent vulnerability of SSL to backdoor attacks. With both empirical and analytical evidence, we reveal that the representation invariance property of SSL, which benefits adversarial robustness, may also be the very reason making SSL highly susceptible to backdoor attacks. Our findings also imply that the existing defenses against supervised backdoor attacks are not easily retrofitted to the unique vulnerability of SSL.
****
Cross-Modal Translation and Alignment for Survival Analysis
Fengtao Zhou, Hao Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21485-21494
With the rapid advances in high-throughput sequencing technologies, the focus of survival analysis has shifted from examining clinical indicators to incorporating genomic profiles with pathological images. However, existing methods either directly adopt a straightforward fusion of pathological features and genomic profiles for survival prediction, or take genomic profiles as guidance to integrate the features of pathological images. The former would overlook intrinsic cross-modal correlations. The latter would discard pathological information irrelevant to gene expression. To address these issues, we present a Cross-Modal Translation and Alignment (CMTA) framework to explore the intrinsic cross-modal correlations and transfer potential complementary information. Specifically, we construct two parallel encoder-decoder structures for multi-modal data to integrate intra-modal information and generate cross-modal representation. Taking the generated cross-modal representation to enhance and recalibrate intra-modal representation can significantly improve its discrimination for comprehensive survival analysis. To explore the intrinsic cross-modal correlations, we further design a cross-modal attention module as the information bridge between different modalities to perform cross-modal interactions and transfer complementary information. Our extensive experiments on five public TCGA datasets demonstrate that our proposed framework outperforms the state-of-the-art methods. The source code has been released.
****
Chaotic World: A Large and Challenging Benchmark for Human Behavior Understanding in Chaotic Events
Kian Eng Ong, Xun Long Ng, Yanchao Li, Wenjie Ai, Kuangyi Zhao, Si Yong Yeo, Jun Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20213-20223
Understanding and analyzing human behaviors (actions and interactions of people), voices, and sounds in chaotic events is crucial in many applications, e.g., crowd management, emergency response services. Different from human behaviors in daily life, human behaviors in chaotic events are generally different in how they behave and influence others, and hence are often much more complex. However, currently there is lack of a large video dataset for analyzing human behaviors in chaotic situations. To this end, we create the first large and challenging multi-modal dataset, Chaotic World, that simultaneously provides different levels of fine-grained and dense spatio-temporal annotations of sounds, individual actions and group interaction graphs, and even text descriptions for each scene in each video, thereby enabling a thorough analysis of complicated behaviors in crowds and chaos. Our dataset consists of a total of 299,923 annotated instances for detecting human behaviors for Spatiotemporal Action Localization in chaotic events, 224,275 instances for identifying interactions between people for Behavior Graph Analysis in chaotic events, 336,390 instances for localizing relevant scenes of interest in long videos for Spatiotemporal Event Grounding, and 378,093 instances for triangulating the source of sound for Event Sound Source Localization. Given the practical complexity and challenges in chaotic events (e.g., large cro

wds, serious occlusions, complicated interaction patterns), our dataset shall be able to facilitate the community to develop, adapt, and evaluate various types of advanced models for analyzing human behaviors in chaotic events. We also design a simple yet effective IntelliCare model with a Dynamic Knowledge Pathfinder module that intelligently learns from multiple tasks and can analyze various aspects of a chaotic scene in a unified architecture. This method achieves promising results in experiments. Dataset and code can be found at https://github.com/sutdcv/Chaotic-World
****
Unsupervised 3D Perception with 2D Vision-Language Distillation for Autonomous Driving
Mahyar Najibi, Jingwei Ji, Yin Zhou, Charles R. Qi, Xinchen Yan, Scott Ettinger, Dragomir Anguelov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8602-8612
Closed-set 3D perception models trained on only a pre-defined set of object categories can be inadequate for safety critical applications such as autonomous driving where new object types can be encountered after deployment. In this paper, we present a multi-modal auto labeling pipeline capable of generating amodal 3D bounding boxes and tracklets for training models on open-set categories without 3D human labels. Our pipeline exploits motion cues inherent in point cloud sequences in combination with the freely available 2D image-text pairs to identify and track all traffic participants. Compared to the recent studies in this domain, which can only provide class-agnostic auto labels limited to moving objects, our method can handle both static and moving objects in the unsupervised manner and is able to output open-vocabulary semantic labels thanks to the proposed vision-language knowledge distillation. Experiments on the Waymo Open Dataset show that our approach outperforms the prior work by significant margins on various unsupervised 3D perception tasks.
****
Towards Grand Unified Representation Learning for Unsupervised Visible-Infrared Person Re-Identification
Bin Yang, Jun Chen, Mang Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11069-11079
Unsupervised learning visible-infrared person re-identification (USL-VI-ReID) is an extremely important and challenging task, which can alleviate the issue of expensive cross-modality annotations. Existing works focus on handling the cross-modality discrepancy under unsupervised conditions. However, they ignore the fact that USL-VI-ReID is a cross-modality retrieval task with the hierarchical discrepancy, i.e., camera variation and modality discrepancy, resulting in clustering inconsistencies and ambiguous cross-modality label association. To address these issues, we propose a hierarchical framework to learn grand unified representation (GUR) for USL-VI-ReID. The grand unified representation lies in two aspects: 1) GUR adopts a bottom-up domain learning strategy with a cross-memory association embedding module to explore the information of hierarchical domains, i.e., intra-camera, inter-camera, and inter-modality domains, learning a unified and robust representation against hierarchical discrepancy. 2) To unify the identities of the two modalities, we develop a cross-modality label unification module that constructs a cross-modality affinity matrix as a bridge for propagating labels between two modalities. Then, we utilize the homogeneous structure matrix to smooth the propagated labels, ensuring that the label structure within one modality remains unchanged. Extensive experiments demonstrate that our GUR framework significantly outperforms existing USL-VI-ReID methods, and even surpasses some supervised counterparts.
****
Active Stereo Without Pattern Projector
Luca Bartolomei, Matteo Poggi, Fabio Tosi, Andrea Conti, Stefano Mattoccia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18470-18482
This paper proposes a novel framework integrating the principles of active stereo in standard passive camera systems without a physical pattern projector. We vi

rtually project a pattern over the left and right images according to the sparse measurements obtained from a depth sensor. Any such devices can be seamlessly plugged into our framework, allowing for the deployment of a virtual active stereo setup in any possible environment, overcoming the limitation of pattern projectors, such as limited working range or environmental conditions. Experiments on indoor/outdoor datasets, featuring both long and close-range, support the seamless effectiveness of our approach, boosting the accuracy of both stereo algorithms and deep networks.
****
Partition Speeds Up Learning Implicit Neural Representations Based on Exponential-Increase Hypothesis
Ke Liu, Feng Liu, Haishuai Wang, Ning Ma, Jiajun Bu, Bo Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5474-5483
Implicit neural representations (INRs) aim to learn a continuous function (i.e., a neural network) to represent an image, where the input and output of the function are pixel coordinates and RGB/Gray values, respectively. However, images tend to consist of many objects whose colors are not perfectly consistent, resulting in the challenge that image is actually a discontinuous piecewise function and cannot be well estimated by a continuous function. In this paper, we empirically investigate that if a neural network is enforced to fit a discontinuous piecewise function to reach a fixed small error, the time costs will increase exponentially with respect to the boundaries in the spatial domain of the target signal. We name this phenomenon the exponential-increase hypothesis. Under the exponential-increase hypothesis, learning INRs for images with many objects will converge very slowly. To address this issue, we first prove that partitioning a complex signal into several sub-regions and utilizing piecewise INRs to fit that signal can significantly speed up the convergence. Based on this fact, we introduce a simple partition mechanism to boost the performance of two INR methods for image reconstruction: one for learning INRs, and the other for learning-to-learn INRs. In both cases, we partition an image into different sub-regions and dedicate smaller networks for each part. In addition, we further propose two partition rules based on regular grids and semantic segmentation maps, respectively. Extensive experiments validate the effectiveness of the proposed partitioning methods in terms of learning INR for a single image (ordinary learning framework) and the learning-to-learn framework.
****
Uncertainty Guided Adaptive Warping for Robust and Efficient Stereo Matching
Junpeng Jing, Jiankun Li, Pengfei Xiong, Jiangyu Liu, Shuaicheng Liu, Yichen Guo, Xin Deng, Mai Xu, Lai Jiang, Leonid Sigal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3318-3327
Correlation based stereo matching has achieved outstanding performance, which pursues cost volume between two feature maps. Unfortunately, current methods with a fixed trained model do not work uniformly well across various datasets, greatly limiting their real-world applicability. To tackle this issue, this paper proposes a new perspective to dynamically calculate correlation for robust stereo matching. A novel Uncertainty Guided Adaptive Correlation (UGAC) module is introduced to robustly adapt the same model for different scenarios. Specifically, a variance-based uncertainty estimation is employed to adaptively adjust the sampling area during warping operation. Additionally, we improve the traditional non-parametric warping with learnable parameters, such that the position-specific weights can be learned. We show that by empowering the recurrent network with the UGAC module, stereo matching can be exploited more robustly and effectively. Extensive experiments demonstrate that our method achieves state-of-the-art performance over the ETH3D, KITTI, and Middlebury datasets when employing the same fixed model over these datasets without any retraining procedure. To target real-time applications, we further design a lightweight model based on UGAC, which also outperforms other methods over KITTI benchmarks with only 0.6 M parameters.
****
ReFit: Recurrent Fitting Network for 3D Human Recovery
Yufu Wang, Kostas Daniilidis; Proceedings of the IEEE/CVF International Conferen

ce on Computer Vision (ICCV), 2023, pp. 14644-14654
We present Recurrent Fitting (ReFit), a neural network architecture for single-i
mage, parametric 3D human reconstruction. ReFit learns a feedback-update loop th
at mirrors the strategy of solving an inverse problem through optimization. At e
ach iterative step, it reprojects keypoints from the human model to feature maps
 to query feedback, and uses a recurrent-based updater to adjust the model to fi
t the image better. Because ReFit encodes strong knowledge of the inverse proble
m, it is faster to train than previous regression models. At the same time, ReFi
t improves state-of-the-art performance on standard benchmarks. Moreover, ReFit
applies to other optimization settings, such as multi-view fitting and single-vi
ew shape fitting. Project website: https://yufu-wang.github.io/refit_humans/
****
Towards Instance-adaptive Inference for Federated Learning
Chun-Mei Feng, Kai Yu, Nian Liu, Xinxing Xu, Salman Khan, Wangmeng Zuo; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 23287-23296
Federated learning (FL) is a distributed learning paradigm that enables multiple
 clients to learn a powerful global model by aggregating local training. However
, the performance of the global model is often hampered by non-i.i.d. distributi
on among the clients, requiring extensive efforts to mitigate inter-client data
heterogeneity. Going beyond inter-client data heterogeneity, we note that intra-
client heterogeneity can also be observed on complex real-world data and serious
ly deteriorate FL performance. In this paper, we present a novel FL algorithm, i
.e., FedIns, to handle intra-client data heterogeneity by enabling instance-adap
tive inference in the FL framework. Instead of huge instance-adaptive models, we
 resort to a parameter-efficient fine-tuning method, i.e., scale and shift deep
features (SSF), upon a pre-trained model. Specifically, we first train an SSF po
ol for each client, and aggregate these SSF pools on the server side, thus still
 maintaining a low communication cost. To enable instance-adaptive inference, fo
r a given instance, we dynamically find the best-matched SSF subsets from the po
ol and aggregate them to generate an adaptive SSF specified for the instance, th
ereby reducing the intra-client as well as the inter-client heterogeneity. Exten
sive experiments show that our FedIns outperforms state-of-the-art FL algorithms
, e.g., a 6.64% improvement against the top-performing method with less than 15%
 communication cost on Tiny-ImageNet.
****
CGBA: Curvature-aware Geometric Black-box Attack
Md Farhamdur Reza, Ali Rahmati, Tianfu Wu, Huaiyu Dai; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2023, pp. 124-133
Decision-based black-box attacks often necessitate a large number of queries to
craft an adversarial example. Moreover, decision-based attacks based on querying
 boundary points in the estimated normal vector direction often suffer from inef
ficiency and convergence issues. In this paper, we propose a novel query-efficie
nt curvature-aware geometric decision-based black-box attack (CGBA) that conduct
s boundary search along a semicircular path on a restricted 2D plane to ensure f
inding a boundary point successfully irrespective of the boundary curvature. Whi
le the proposed CGBA attack can work effectively for an arbitrary decision bound
ary, it is particularly efficient in exploiting the low curvature to craft high-
quality adversarial examples, which is widely seen and experimentally verified i
n commonly used classifiers under non-targeted attacks. In contrast, the decisio
n boundaries often exhibit higher curvature under targeted attacks. Thus, we dev
elop a new query-efficient variant, CGBA-H, that is adapted for the targeted att
ack. In addition, we further design an algorithm to obtain a better initial boun
dary point at the expense of some extra queries, which considerably enhances the
 performance of the targeted attack. Extensive experiments are conducted to eval
uate the performance of our proposed methods against some well-known classifiers
 on the ImageNet and CIFAR10 datasets, demonstrating the superiority of CGBA and
 CGBA-H over state-of-the-art non-targeted and targeted attacks, respectively. T
he source code is available at https://github.com/Farhamdur/CGBA.
****

Unsupervised Facial Performance Editing via Vector-Quantized StyleGAN Representa
tions
Berkay Kicanaoglu, Pablo Garrido, Gaurav Bharaj; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 2371-2382
High-fidelity virtual human avatar applications create a need for photorealistic
 video face synthesis with controllable semantic editing over facial features. W
hile recent generative neural methods have shown significant progress in portrai
t video synthesis, intuitive facial control, e.g., of mouth interior and gaze at
 different levels of details, remains a challenge. In this work, we present a no
vel face editing framework that combines a 3D face model with StyleGAN vector-qu
antization to learn multi-level semantic facial control. We show that vector qua
ntization of StyleGAN features unveils richer semantic facial representations, e
.g., teeth and pupils, which are difficult to model with 3D tracking priors. Suc
h representations along with 3D tracking can be used as self-supervision to trai
n a generator with control over coarse expressions and finer facial attributes.
Learned representations can be combined with user-defined masks to create semant
ic segmentations that act as custom detail handles for semantic-aware video edit
ing. Our formulation allows video face manipulation with precise local control o
ver facial attributes, such as eyes and teeth, opening up a number of face reena
ctment and visual expression articulation applications.
****
Online Clustered Codebook
Chuanxia Zheng, Andrea Vedaldi; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 22798-22807
Vector Quantisation (VQ) is experiencing a comeback in machine learning, where i
t is increasingly used in representation learning. However, optimizing the codev
ectors in existing VQ-VAE is not entirely trivial. A problem is codebook collaps
e, where only a small subset of codevectors receive gradients useful for their o
ptimization, whereas a majority of them simply "dies off" and is never updated o
r used. This limits the effectiveness of VQ for learning larger codebooks in com
plex computer vision tasks that require high-capacity representations. In this p
aper, we present a simple alternative method for online codebook learning, Clust
ering VQ-VAE (CVQ-VAE).
 Our approach selects encoded features as anchors to update the "dead" codevecto
rs, while optimizing the codebooks which are alive via the original loss. This s
trategy brings unused codevectors closer in distribution to the encoded features
, increasing the likelihood of being chosen and optimized. We extensively valida
te the generalization capability of our quantizer on various datasets, tasks (e.
g., reconstruction and generation), and architectures (e.g., VQ-VAE, VQGAN, LDM)
. CVQ-VAE can be easily integrated into the existing models with just a few line
s of code.
****
A Multidimensional Analysis of Social Biases in Vision Transformers
Jannik Brinkmann, Paul Swoboda, Christian Bartelt; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 4914-4923
The embedding spaces of image models have been shown to encode a range of social
 biases such as racism and sexism. Here, we investigate specific factors that co
ntribute to the emergence of these biases in Vision Transformers (ViT). Therefor
e, we measure the impact of training data, model architecture, and training obje
ctives on social biases in the learned representations of ViTs. Our findings ind
icate that counterfactual augmentation training using diffusion-based image edit
ing can mitigate biases, but does not eliminate them. Moreover, we find that lar
ger models are less biased than smaller models, and that models trained using di
scriminative objectives are less biased than those trained using generative obje
ctives. In addition, we observe inconsistencies in the learned social biases. To
 our surprise, ViTs can exhibit opposite biases when trained on the same data se
t using different self-supervised objectives. Our findings give insights into th
e factors that contribute to the emergence of social biases and suggests that we
 could achieve substantial fairness improvements based on model design choices.
****

PGFed: Personalize Each Client's Global Objective for Federated Learning
Jun Luo, Matias Mendieta, Chen Chen, Shandong Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3946-3956
Personalized federated learning has received an upsurge of attention due to the mediocre performance of conventional federated learning (FL) over heterogeneous data. Unlike conventional FL which trains a single global consensus model, personalized FL allows different models for different clients. However, existing personalized FL algorithms only implicitly transfer the collaborative knowledge across the federation by embedding the knowledge into the aggregated model or regularization. We observed that this implicit knowledge transfer fails to maximize the potential of each client's empirical risk toward other clients. Based on our observation, in this work, we propose Personalized Global Federated Learning (PGFed), a novel personalized FL framework that enables each client to personalize its own global objective by explicitly and adaptively aggregating the empirical risks of itself and other clients. To avoid massive (O(N^2)) communication overhead and potential privacy leakage while achieving this, each client's risk is estimated through a first-order approximation for other clients' adaptive risk aggregation. On top of PGFed, we develop a momentum upgrade, dubbed PGFedMo, to more efficiently utilize clients' empirical risks. Our extensive experiments on four datasets under different federated settings show consistent improvements of PGFed over previous state-of-the-art methods. The code is publicly available at https://github.com/ljaiverson/pgfed.
****
Verbs in Action: Improving Verb Understanding in Video-Language Models
Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15579-15591
Understanding verbs is crucial to modelling how people and objects interact with each other and the environment through space and time. Recently, state-of-the-art video-language models based on CLIP have been shown to have limited verb understanding and to rely extensively on nouns, restricting their performance in real-world video applications that require action and temporal understanding. In this work, we improve verb understanding for CLIP-based video-language models by proposing a new Verb-Focused Contrastive (VFC) framework. This consists of two main components: (1) leveraging pretrained large language models (LLMs) to create hard negatives for cross-modal contrastive learning, together with a calibration strategy to balance the occurrence of concepts in positive and negative pairs; and (2) enforcing a fine-grained, verb phrase alignment loss. Our method achieves state-of-the-art results for zero-shot performance on three downstream tasks that focus on verb understanding, including video-text matching, video question-answering and video classification; while maintaining performance on noun-focused settings. To the best of our knowledge, this is the first work which proposes a method to alleviate the verb understanding problem, and does not simply highlight it. Our code is publicly available.
****
Zero-Shot Point Cloud Segmentation by Semantic-Visual Aware Synthesis
Yuwei Yang, Munawar Hayat, Zhao Jin, Hongyuan Zhu, Yinjie Lei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11586-11596
This paper proposes a feature synthesis approach for zero-shot semantic segmentation of 3D point clouds, enabling generalization to previously unseen categories. Given only the class-level semantic information for unseen objects, we strive to enhance the correspondence, alignment and consistency between the visual and semantic spaces, to synthesise diverse, generic and transferable visual features. We develop a masked learning strategy to promote diversity within the same class visual features and enhance the separation between different classes. We further cast the visual features into a prototypical space to model their distribution for alignment with the corresponding semantic space. Finally, we develop a consistency regularizer to preserve the semantic-visual relationships between the real-seen features and synthetic-unseen features. Our approach shows considerabl

e semantic segmentation gains on ScanNet, S3DIS and SemanticKITTI benchmarks. Ou
r code is available at: https://github.com/leolyj/3DPC-GZSL
****
Exploring Predicate Visual Context in Detecting of Human-Object Interactions
Frederic Z Zhang, Yuhui Yuan, Dylan Campbell, Zhuoyao Zhong, Stephen Gould; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023
, pp. 10411-10421
Recently, the DETR framework has emerged as the dominant approach for human--obj
ect interaction (HOI) research. In particular, two-stage transformer-based HOI d
etectors are amongst the most performant and training-efficient approaches. Howe
ver, these often condition HOI classification on object features that lack fine-
grained contextual information, eschewing pose and orientation information in fa
vour of visual cues about object identity and box extremities. This naturally hi
nders the recognition of complex or ambiguous interactions. In this work, we stu
dy these issues through visualisations and carefully designed experiments. Accor
dingly, we investigate how best to re-introduce image features via cross-attenti
on. With an improved query design, extensive exploration of keys and values, and
 box pair positional embeddings as spatial guidance, our model with enhanced pre
dicate visual context (PViC) outperforms state-of-the-art methods on the HICO-DE
T and V-COCO benchmarks, while maintaining low training cost.
****
Robo3D: Towards Robust and Reliable 3D Perception against Corruptions
Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang
 Pan, Kai Chen, Ziwei Liu; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2023, pp. 19994-20006
The robustness of 3D perception systems under natural corruptions from environme
nts and sensors is pivotal for safety-critical applications. Existing large-scal
e 3D perception datasets often contain data that are meticulously cleaned. Such
configurations, however, cannot reflect the reliability of perception models dur
ing the deployment stage. In this work, we present Robo3D, the first comprehensi
ve benchmark heading toward probing the robustness of 3D detectors and segmentor
s under out-of-distribution scenarios against natural corruptions that occur in
real-world environments. Specifically, we consider eight corruption types stemmi
ng from severe weather conditions, external disturbances, and internal sensor fa
ilure. We uncover that, although promising results have been progressively achie
ved on standard benchmarks, state-of-the-art 3D perception models are at risk of
 being vulnerable to corruptions. We draw key observations on the use of data re
presentations, augmentation schemes, and training strategies, that could severel
y affect the model's performance. To pursue better robustness, we propose a dens
ity-insensitive training framework along with a simple flexible voxelization str
ategy to enhance the model resiliency. We hope our benchmark and approach could
inspire future research in designing more robust and reliable 3D perception mode
ls. Our robustness benchmark suite is publicly available.
****
Towards Saner Deep Image Registration
Bin Duan, Ming Zhong, Yan Yan; Proceedings of the IEEE/CVF International Confere
nce on Computer Vision (ICCV), 2023, pp. 12459-12468
With recent advances in computing hardware and surges of deep-learning architect
ures, learning-based deep image registration methods have surpassed their tradit
ional counterparts, in terms of metric performance and inference time. However,
these methods focus on improving performance measurements such as Dice, resultin
g in less attention given to model behaviors that are equally desirable for regi
strations, especially for medical imaging. This paper investigates these behavio
rs for popular learning-based deep registrations under a sanity-checking microsc
ope. We find that most existing registrations suffer from low inverse consistenc
y and nondiscrimination of identical pairs due to overly optimized image similar
ities. To rectify these behaviors, we propose a novel regularization-based sanit
y-enforcer method that imposes two sanity checks on the deep model to reduce its
 inverse consistency errors and increase its discriminative power simultaneously
. Moreover, we derive a set of theoretical guarantees for our sanity-checked ima

ge registration method, with experimental results supporting our theoretical findings and their effectiveness in increasing the sanity of models without sacrificing any performance.
****
Instance and Category Supervision are Alternate Learners for Continual Learning
Xudong Tian, Zhizhong Zhang, Xin Tan, Jun Liu, Chengjie Wang, Yanyun Qu, Guannan Jiang, Yuan Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5596-5605
Continual Learning (CL) is the constant development of complex behaviors by building upon previously acquired skills. Yet, current CL algorithms tend to incur class-level forgetting as the label information is often quickly overwritten by new knowledge. This motivates attempts to mine instance-level discrimination by resorting to recent self-supervised learning (SSL) techniques. However, previous works have pointed that the self-supervised learning objective is essentially a trade-off between invariance to distortion and preserving sample information, which seriously hinders the unleashing of instance-level discrimination.

In this work, we reformulate SSL from the information-theoretic perspective by disentangling the goal of instance-level discrimination, and tackle the trade-off to promote compact representations with maximally preserved invariance to distortion. On this basis, we develop a novel alternate learning paradigm to enjoy the complementary merits of instance-level and category-level supervision, which yields improved robustness against forgetting and better adaptation to each task. To verify the proposed method, we conduct extensive experiments on four different benchmarks using both class-incremental and task-incremental settings, where the leap in performance and thorough ablation studies demonstrate the efficacy and efficiency of our modeling strategy.
****
Diverse Data Augmentation with Diffusions for Effective Test-time Prompt Tuning
Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, Wangmeng Zuo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2704-2714
Benefiting from prompt tuning, recent years have witnessed the promising performance of pre-trained vision-language models, e.g., CLIP, on versatile downstream tasks. In this paper, we focus on a particular setting of learning adaptive prompts on the fly for each test sample from an unseen new domain, which is known as test-time prompt tuning (TPT). Existing TPT methods typically rely on data augmentation and confidence selection. However, conventional data augmentation techniques, e.g., random resized crops, suffers from the lack of data diversity, while entropy-based confidence selection alone is not sufficient to guarantee prediction fidelity. To address these issues, we propose a novel TPT method, named DiffTPT, which leverages pre-trained diffusion models to generate diverse and informative new data. Specifically, we incorporate augmented data by both conventional method and pre-trained stable diffusion to exploit their respective merits, improving the model's ability to adapt to unknown new test data. Moreover, to ensure the prediction fidelity of generated data, we introduce a cosine similarity-based filtration technique to select the generated data with higher similarity to the single test sample. Our experiments on test datasets with distribution shifts and unseen categories demonstrate that DiffTPT improves the zero-shot accuracy by an average of 5.13% compared to the state-of-the-art TPT method.
****
Interaction-aware Joint Attention Estimation Using People Attributes
Chihiro Nakatani, Hiroaki Kawashima, Norimichi Ukita; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10224-10233
This paper proposes joint attention estimation in a single image. Different from related work in which only the gaze-related attributes of people are independently employed, (I) their locations and actions are also employed as contextual cues for weighting their attributes, and (ii) interactions among all of these attributes are explicitly modeled in our method. For the interaction modeling, we propose a novel Transformer-based attention network to encode joint attention as low-dimensional features. We introduce a specialized MLP head with positional emb

edding to the Transformer so that it predicts pixelwise confidence of joint atte ntion for generating the confidence heatmap. This pixelwise prediction improves the heatmap accuracy by avoiding the ill-posed problem in which the high-dimensi onal heatmap is predicted from the low-dimensional features. The estimated joint attention is further improved by being integrated with general image-based atte ntion estimation. Our method outperforms SOTA methods quantitatively in comparat ive experiments. Code: https://anonymous.4open.science/r/anonymized_codes-ECA4.
****
GePSAn: Generative Procedure Step Anticipation in Cooking Videos
Mohamed A. Abdelsalam, Samrudhdhi B. Rangrej, Isma Hadji, Nikita Dvornik, Konsta ntinos G. Derpanis, Afsaneh Fazly; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2023, pp. 2988-2997
We study the problem of future step anticipation in procedural videos. Given a v ideo of an ongoing procedural activity, we predict a plausible next procedure st ep described in rich natural language. While most previous work focus on the pro blem of data scarcity in procedural video datasets, another core challenge of fu ture anticipation is how to account for multiple plausible future realizations i n natural settings. This problem has been largely overlooked in previous work. T o address this challenge, we frame future step prediction as modelling the distr ibution of all possible candidates for the next step. Specifically, we design a generative model that takes a series of video clips as input, and generates mult iple plausible and diverse candidates (in natural language) for the next step. F ollowing previous work, we side-step the video annotation scarcity by pretrainin g our model on a large text-based corpus of procedural activities, and then tran sfer the model to the video domain. Our experiments, both in textual and video d omains, show that our model captures diversity in the next step prediction and g enerates multiple plausible future predictions. Moreover, our model establishes new state-of-the-art results on YouCookII, where it outperforms existing baselin es on the next step anticipation. Finally, we also show that our model can succe ssfully transfer from text to the video domain zero-shot, ie, without fine-tunin g or adaptation, and produces good-quality future step predictions from video.
****
Gradient-based Sampling for Class Imbalanced Semi-supervised Object Detection
Jiaming Li, Xiangru Lin, Wei Zhang, Xiao Tan, Yingying Li, Junyu Han, Errui Ding , Jingdong Wang, Guanbin Li; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2023, pp. 16390-16400
Current semi-supervised object detection (SSOD) algorithms typically assume clas s balanced datasets (PASCAL VOC etc.) or slightly class imbalanced datasets (MSC OCO, etc). This assumption can be easily violated since real world datasets can be extremely class imbalanced in nature, thus making the performance of semi-sup ervised object detectors far from satisfactory. Besides, the research for this p roblem in SSOD is severely under-explored. To bridge this research gap, we compr ehensively study the class imbalance problem for SSOD under more challenging sce narios, thus forming the first experimental setting for class imbalanced SSOD (C I-SSOD). Moreover, we propose a simple yet effective gradient-based sampling fra mework that tackles the class imbalance problem from the perspective of two type s of confirmation biases. To tackle confirmation bias towards majority classes, the gradient-based reweighting and gradient-based thresholding modules leverage the gradients from each class to fully balance
the influence of the majority and minority classes. To tackle the confirmation bias from incorrect pseudo labels of minority classes, the class-rebalancing sam pling module resamples unlabeled data following the guidance of the gradient-bas ed reweighting module. Experiments on three proposed sub-tasks, namely MS-COCO, MS-COCO- Object365 and LVIS, suggest that our method outperforms
current class imbalanced object detectors by clear margins, serving as a baseli ne for future research in CISSOD. Code will be available at https://github.com/n ightkeepers/CI-SSOD.
****
SLCA: Slow Learner with Classifier Alignment for Continual Learning on a Pre-tra ined Model

Gengwei Zhang, Liyuan Wang, Guoliang Kang, Ling Chen, Yunchao Wei; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 191
48-19158
The goal of continual learning is to improve the performance of recognition mode
ls in learning sequentially arrived data. Although most existing works are estab
lished on the premise of learning from scratch, growing efforts have been devote
d to incorporating the benefits of pre-training. However, how to adaptively expl
oit the pre-trained knowledge for each incremental task while maintaining its ge
neralizability remains an open question. In this work, we present an extensive a
nalysis for continual learning on a pre-trained model (CLPM), and attribute the
key challenge to a progressive overfitting problem. Observing that selectively r
educing the learning rate can almost resolve this issue in the representation la
yer, we propose a simple but extremely effective approach named Slow Learner wit
h Classifier Alignment (SLCA), which further improves the classification layer b
y modeling the class-wise distributions and aligning the classification layers i
n a post-hoc fashion. Across a variety of scenarios, our proposal provides subst
antial improvements for CLPM (e.g., up to 49.76%, 50.05%, 44.69% and 40.16% on S
plit CIFAR-100, Split ImageNet-R, Split CUB-200 and Split Cars-196, respectively
), and thus outperforms state-of-the-art approaches by a large margin. Based on
such a strong baseline, critical factors and promising directions are analyzed i
n-depth to facilitate subsequent research.
****
Implicit Temporal Modeling with Learnable Alignment for Video Recognition
Shuyuan Tu, Qi Dai, Zuxuan Wu, Zhi-Qi Cheng, Han Hu, Yu-Gang Jiang; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19
936-19947
Contrastive language-image pretraining (CLIP) has demonstrated remarkable succes
s in various image tasks. However, how to extend CLIP with effective temporal mo
deling is still an open and crucial problem. Existing factorized or joint spatia
l-temporal modeling trades off between the efficiency and performance. While mod
eling temporal information within straight through tube is widely adopted in lit
erature, we find that simple frame alignment already provides enough essence wit
hout temporal attention. To this end, in this paper, we proposed a novel Implici
t Learnable Alignment (ILA) method, which minimizes the temporal modeling effort
while achieving incredibly high performance. Specifically, for a frame pair, an
interactive point is predicted in each frame, serving as the mutual information
rich region. By enhancing the features around the interactive point, two frames
are implicitly aligned. The aligned features are then pooled into a single toke
n, which is leveraged in the subsequent spatial self-attention. Our method allow
s eliminating the costly or insufficient temporal self-attention in video. Exten
sive experiments on benchmarks demonstrate the superiority and generality of our
module. Particularly, the proposed ILA achieves a top-1 accuracy of 88.9% on Ki
netics-400 with much fewer FLOPs compared with Swin-L and ViViT-H. Code is relea
sed at https://github.com/Francis-Rings/ILA.
****
Non-Coaxial Event-Guided Motion Deblurring with Spatial Alignment
Hoonhee Cho, Yuhwan Jeong, Taewoo Kim, Kuk-Jin Yoon; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 12492-12503
Motion deblurring from a blurred image is a challenging computer vision problem
because frame-based cameras lose information during the blurring process. Severa
l attempts have compensated for the loss of motion information by using event ca
meras, which are bio-inspired sensors with a high temporal resolution. Even thou
gh most studies have assumed that image and event data are pixel-wise aligned, t
his is only possible with low-quality active-pixel sensor (APS) images and synth
etic datasets. In real scenarios, obtaining per-pixel aligned event-RGB data is
technically challenging since event and frame cameras have different optical axe
s. For the application of the event camera, we propose the first Non-coaxial Eve
nt-guided Image Deblurring (NEID) approach that utilizes the camera setup compos
ed of a standard frame-based camera with a non-coaxial single event camera. To c
onsider the per-pixel alignment between the image and event without additional d

evices, we propose the first NEID network that spatially aligns events to images while refining the image features from temporally dense event features. For training and evaluation of our network, we also present the first large-scale dataset, consisting of RGB frames with non-aligned events aimed at a breakthrough in motion deblurring with an event camera. Extensive experiments on various datasets demonstrate that the proposed method achieves significantly better results than the prior works in terms of performance and speed, and it can be applied for practical uses of event cameras.
****
Fingerprinting Deep Image Restoration Models
Yuhui Quan, Huan Teng, Ruotao Xu, Jun Huang, Hui Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13285-13295
Fingerprinting is a promising non-invasive method for protecting the intellectual property rights (IPR) of deep neural network (DNN) models. It extracts a feature called a fingerprint from a DNN model to identify its ownership. Existing fingerprinting methods focus only on classification-related models that map images to labels, while inapplicable to models for image restoration that map images to images. This paper proposes a fingerprinting framework for DNN models of image restoration. The proposed framework defines the fingerprint using a critical image, which exhibits strongly discriminative patterns and is robust to modest model modifications. Model ownership is then verified by comparing the distance of color histograms and local gradient pattern histograms of critical images between the suspect and source models. We apply the proposed framework to two representative tasks, denoising and super-resolution. It outperforms the baselines of fingerprinting and competes against existing invasive model watermarking methods.
****
AutoDiffusion: Training-Free Optimization of Time Steps and Architectures for Automated Diffusion Model Acceleration
Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7105-7114
Diffusion models are emerging expressive generative models, in which a large number of time steps (inference steps) are required for a single image generation. To accelerate such tedious process, reducing steps uniformly is considered as an undisputed principle of diffusion models. We consider that such a uniform assumption is not the optimal solution in practice; i.e., we can find different optimal time steps for different models. Therefore, we propose to search the optimal time steps sequence and compressed model architecture in a unified framework to achieve effective image generation for diffusion models without any further training. Specifically, we first design a unified search space that consists of all possible time steps and various architectures. Then, a two stage evolutionary algorithm is introduced to find the optimal solution in the designed search space. To further accelerate the search process, we employ FID score between generated and real samples to estimate the performance of the sampled examples. As a result, the proposed method is (i).training-free, obtaining the optimal time steps and model architecture without any training process; (ii). orthogonal to most advanced diffusion samplers and can be integrated to gain better sample quality. (iii). generalized, where the searched time steps and architectures can be directly applied on different diffusion models with the same guidance scale. Experimental results show that our method achieves excellent performance by using only a few time steps, e.g. 17.86 FID score on ImageNet 64 x 64 with only four steps, compared to 138.66 with DDIM.
****
SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes
Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, Limin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9921-9931
Multi-object tracking (MOT) in sports scenes plays a critical role in gathering players statistics, supporting further applications, such as automatic tactical analysis. Yet existing MOT benchmarks cast little attention on this domain. In t

his work, we present a new large-scale multi-object tracking dataset in multiple
 sports scenes, coined as SportsMOT, where all players on the court are supposed
 to be tracked. It consists of 240 video sequences, over 150K frames (almost 15x
 MOT17) and over 1.6M bounding boxes (3x MOT17) collected from 3 sports categori
es, including basketball, volleyball and football. Our dataset is characterized
with two key properties: 1) fast and variable-speed motion and 2) similar yet di
stinguishable appearance. We expect SportsMOT to encourage the MOT trackers to p
romote in both motion-based association and appearance-based association. We ben
chmark several state-of-the-art trackers and reveal the key challenge of SportsM
OT lies in object association. To alleviate the issue, we further propose a new
multi-object tracking framework, termed as MixSort, introducing a MixFormer-like
 structure as an auxiliary association model to prevailing tracking-by-detection
 trackers. By integrating the customized appearance-based association with the o
riginal motion-based association, MixSort achieves state-of-the-art performance
on SportsMOT and MOT17. Based on MixSort, we give an in-depth analysis and provi
de some profound insights into SportsMOT.
****
Localizing Moments in Long Video Via Multimodal Guidance
Wayner Barrios, Mattia Soldan, Alberto Mario Ceballos-Arroyo, Fabian Caba Heilbr
on, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Comp
uter Vision (ICCV), 2023, pp. 13667-13678
The recent introduction of the large-scale, long-form MAD and Ego4D datasets has
 enabled researchers to investigate the performance of current state-of-the-art
methods for video grounding in the long-form setup, with interesting findings: c
urrent grounding methods alone fail at tackling this challenging task and setup
due to their inability to process long video sequences. In this paper, we propos
e a method for improving the performance of natural language grounding in long v
ideos by identifying and pruning out non-describable windows. We design a guided
 grounding framework consisting of a Guidance Model and a base grounding model.
The Guidance Model emphasizes describable windows, while the base grounding mode
l analyzes short temporal windows to determine which segments accurately match a
 given language query. We offer two designs for the Guidance Model: Query-Agnost
ic and Query-Dependent, which balance efficiency and accuracy. Experiments demon
strate that our proposed method outperforms state-of-the-art models by 4.1% in M
AD and 4.52% in Ego4D (NLQ), respectively. Code, data and MAD's audio features n
ecessary to reproduce our experiments are available at: https://github.com/wayba
rrios/guidance-based-video-grounding.
****
Pixel-Aligned Recurrent Queries for Multi-View 3D Object Detection
Yiming Xie, Huaizu Jiang, Georgia Gkioxari, Julian Straub; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18370-18380
We present PARQ - a multi-view 3D object detector with transformer and pixel-ali
gned recurrent queries. Unlike previous works that use learnable features or onl
y encode 3D point positions as queries in the decoder, PARQ leverages appearance
-enhanced queries initialized from reference points in 3D space and updates thei
r 3D location with recurrent cross-attention operations. Incorporating pixel-ali
gned features and cross attention enables the model to encode the necessary 3D-t
o-2D correspondences and capture global contextual information of the input imag
es. PARQ outperforms prior best methods on the ScanNet and ARKitScenes datasets,
 learns and detects faster, is more robust to distribution shifts in reference p
oints, can leverage additional input views without retraining, and can adapt inf
erence compute by changing the number of recurrent iterations.
****
Towards Universal Image Embeddings: A Large-Scale Dataset and Challenge for Gene
ric Image Representations
Nikolaos-Antonios Ypsilantis, Kaifeng Chen, Bingyi Cao, Mário Lipovský, Pelin Do
gan-Schönberger, Grzegorz Makosa, Boris Bluntschli, Mojtaba Seyedhosseini, Ond■e
j Chum, André Araujo; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2023, pp. 11290-11301
Fine-grained and instance-level recognition methods are commonly trained and eva

luated on specific domains, in a model per domain scenario. Such an approach, however, is impractical in real large-scale applications. In this work, we address the problem of universal image embedding, where a single universal model is trained and used in multiple domains. First, we leverage existing domain-specific datasets to carefully construct a new large-scale public benchmark for the evaluation of universal image embeddings, with 241k query images, 1.4M index images and 2.8M training images across 8 different domains and 349k classes.
We define suitable metrics, training and evaluation protocols to foster future research in this area. Second, we provide a comprehensive experimental evaluation on the new dataset, demonstrating that existing approaches and simplistic extensions lead to worse performance than an assembly of models trained for each domain separately. Finally, we conducted a public research competition on this topic, leveraging industrial datasets, which attracted the participation of more than 1k teams worldwide. This exercise generated many interesting research ideas and findings which we present in detail. Project webpage: https://cmp.felk.cvut.cz/univ_emb/
****
SemARFlow: Injecting Semantics into Unsupervised Optical Flow Estimation for Autonomous Driving
Shuai Yuan, Shuzhi Yu, Hannah Kim, Carlo Tomasi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9566-9577
Unsupervised optical flow estimation is especially hard near occlusions and motion boundaries and in low-texture regions. We show that additional information such as semantics and domain knowledge can help better constrain this problem. We introduce SemARFlow, an unsupervised optical flow network designed for autonomous driving data that takes estimated semantic segmentation masks as additional inputs. This additional information is injected into the encoder and into a learned upsampler that refines the flow output. In addition, a simple yet effective semantic augmentation module provides self-supervision when learning flow and its boundaries for vehicles, poles, and sky. Together, these injections of semantic information improve the KITTI-2015 optical flow test error rate from 11.80% to 8.38%. We also show visible improvements around object boundaries as well as a greater ability to generalize across datasets. Code is available at https://github.com/duke-vision/semantic-unsup-flow-release.
****
TiDAL: Learning Training Dynamics for Active Learning
Seong Min Kye, Kwanghee Choi, Hyeongmin Byun, Buru Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22335-22345
Active learning (AL) aims to select the most useful data samples from an unlabeled data pool and annotate them to expand the labeled dataset under a limited budget.Especially, uncertainty-based methods choose the most uncertain samples, which are known to be effective in improving model performance.However, AL literature often overlooks training dynamics (TD), defined as the ever-changing model behavior during optimization via stochastic gradient descent, even though other research areas have empirically shown that TD provides important clues for measuring the data uncertainty. In this paper, we first provide theoretical and empirical evidence to argue the usefulness of utilizing the ever-changing model behavior rather than the fully trained model snapshot. We then propose a novel AL method, Training Dynamics for Active Learning (TiDAL), which efficiently predicts the training dynamics of unlabeled data to estimate their uncertainty. Experimental results show that our TiDAL achieves better or comparable performance on both balanced and imbalanced benchmark datasets compared to state-of-the-art AL methods, which estimate data uncertainty using only static information after model training.
****
Uncertainty-aware Unsupervised Multi-Object Tracking
Kai Liu, Sheng Jin, Zhihang Fu, Ze Chen, Rongxin Jiang, Jieping Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9996-10005
Without manually annotated identities, unsupervised multi-object trackers are in

ferior to learning reliable feature embeddings. It causes the similarity-based inter-frame association stage also be error-prone, where an uncertainty problem arises. The frame-by-frame accumulated uncertainty prevents trackers from learning the consistent feature embedding against time variation. To avoid this uncertainty problem, recent self-supervised techniques are adopted, whereas they failed to capture temporal relations. The inter-frame uncertainty still exists. In fact, this paper argues that though the uncertainty problem is inevitable, it is possible to leverage the uncertainty itself to improve the learned consistency in turn. Specifically, an uncertainty-based metric is developed to verify and rectify the risky associations. The resulting accurate pseudo-tracklets boost learning the feature consistency. And accurate tracklets can incorporate temporal information into spatial transformation. This paper proposes a tracklet-guided augmentation strategy to simulate the tracklet's motion, which adopts a hierarchical uncertainty-based sampling mechanism for hard sample mining. The ultimate unsupervised MOT framework, namely U2MOT, is proven effective on MOT-Challenges and VisDrone-MOT benchmark. U2MOT achieves a SOTA performance among the published supervised and unsupervised trackers.
****
DPS-Net: Deep Polarimetric Stereo Depth Estimation
Chaoran Tian, Weihong Pan, Zimo Wang, Mao Mao, Guofeng Zhang, Hujun Bao, Ping Tan, Zhaopeng Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3569-3579
Stereo depth estimation usually struggles to deal with textureless scenes for both traditional and learning-based methods due to the inherent dependence on image correspondence matching. In this paper, we propose a novel neural network, i.e., DPS-Net, to exploit both the prior geometric knowledge and polarimetric information for depth estimation with two polarimetric stereo images. Specifically, we construct both RGB and polarization correlation volumes to fully leverage the multi-domain similarity between polarimetric stereo images. Since inherent ambiguities exist in the polarization images, we introduce the iso-depth cost explicitly into the network to solve these ambiguities. Moreover, we design a cascaded dual-GRU architecture to recurrently update the disparity and effectively fuse both the multi-domain correlation features and the iso-depth cost. Besides, we present new synthetic and real polarimetric stereo datasets for evaluation. Experimental results demonstrate that our method outperforms the state-of-the-art stereo depth estimation methods.
****
Designing Phase Masks for Under-Display Cameras
Anqi Yang, Eunhee Kang, Hyong-Euk Lee, Aswin C. Sankaranarayanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10637-10645
Diffractive blur and low light levels are two fundamental challenges in producing high-quality photographs in under-display cameras (UDCs).
In this paper, we incorporate phase masks on display panels to tackle both challenges. Our design inserts two phase masks, specifically two microlens arrays, in front of and behind a display panel. The first phase mask concentrates light on the locations where the display is transparent so that more light passes through the display, and the second phase mask reverts the effect of the first phase mask. We further optimize the folding height of each microlens to improve the quality of PSFs and suppress chromatic aberration. We evaluate our design using a physically-accurate simulator based on Fourier optics. The proposed design is able to double the light throughput while improving the invertibility of the PSFs. Lastly, we discuss the effect of our design on the display quality and show that implementation with polarization-dependent phase masks can leave the display quality uncompromised.
****
Can Language Models Learn to Listen?
Evonne Ng, Sanjay Subramanian, Dan Klein, Angjoo Kanazawa, Trevor Darrell, Shiry Ginosar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10083-10093

We present a framework for generating appropriate facial responses from a listener in dyadic social interactions based on the speaker's words. Given an input transcription of the speaker's words with their timestamps, our approach autoregressively predicts a response of a listener: a sequence of listener facial gestures, quantized using a VQ-VAE. Since gesture is a language component, we propose treating the quantized atomic motion elements as additional language token inputs to a transformer-based large language model. Initializing our transformer with the weights of a language model pre-trained only on text results in significantly higher quality listener responses than training a transformer from scratch. We show that our generated listener motion is fluent and reflective of language semantics through quantitative metrics and a qualitative user study. In our evaluation, we analyze the model's ability to utilize temporal and semantic aspects of spoken text.
****
SpaceEvo: Hardware-Friendly Search Space Design for Efficient INT8 Inference
Xudong Wang, Li Lyna Zhang, Jiahang Xu, Quanlu Zhang, Yujing Wang, Yuqing Yang, Ningxin Zheng, Ting Cao, Mao Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5819-5828
The combination of Neural Architecture Search (NAS) and quantization has proven successful in automatically designing low-FLOPs INT8 quantized neural networks (QNN). However, directly applying NAS to design accurate QNN models that achieve low latency on real-world devices leads to inferior performance. In this work, we identify that the poor INT8 latency is due to the quantization-unfriendly issue: the operator and configuration (e.g., channel width) choices in prior art search spaces lead to diverse quantization efficiency and can slow down the INT8 inference speed. To address this challenge, we propose SpaceEvo, an automatic method for designing a dedicated, quantization-friendly search space for each target hardware. The key idea of SpaceEvo is to automatically search hardware-preferred operators and configurations to construct the search space, guided by a metric called Q-T score to quantify how quantization-friendly a candidate search space is. We further train a quantized-for-all supernet over our discovered search space, enabling the searched models to be directly deployed without extra retraining or quantization. Our discovered models, SEQnet, establish new SOTA INT8 quantized accuracy under various latency constraints, achieving up to 10.1% accuracy improvement on ImageNet than prior art CNNs under the same latency. Extensive experiments on real devices show that SpaceEvo consistently outperforms manually-designed search spaces with up to 2.5x faster speed while achieving the same accuracy.
****
How Far Pre-trained Models Are from Neural Collapse on the Target Dataset Informs their Transferability
Zijian Wang, Yadan Luo, Liang Zheng, Zi Huang, Mahsa Baktashmotlagh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5549-5558
This paper focuses on model transferability estimation, i.e., assessing the performance of pre-trained models on a downstream task without performing fine-tuning. Motivated by the neural collapse (NC) that reveals the feature geometry at the terminal stage of training, our method considers the model transferability as how far the target activations obtained by pre-trained models are from their hypothetical state in the terminal phase of the fine-tuned model. We propose a metric that computes this proximity based on three phenomena of NC: within-class variability collapse, simplex encoded label interpolation geometry structure is formed, and the nearest center classifier becomes optimal on training data. Extensive experiments on 11 benchmark datasets demonstrate the effectiveness and efficiency of the proposed method over the existing SOTA approaches. Particularly, our method achieves SOTA transferability estimation accuracy with approximately 10x wall-clock time speed up compared to the existing approaches
****
SurfsUP: Learning Fluid Simulation for Novel Surfaces
Arjun Mani, Ishaan Preetam Chandratreya, Elliot Creager, Carl Vondrick, Richard

Zemel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14225-14235
Modeling the mechanics of fluid in complex scenes is vital to applications in design, graphics, and robotics. Learning-based methods provide fast and differentiable fluid simulators, however most prior work is unable to accurately model how fluids interact with genuinely novel surfaces not seen during training. We introduce SurfsUP, a framework that represents objects implicitly using signed distance functions (SDFs), rather than an explicit representation of meshes or particles. This continuous representation of geometry enables more accurate simulation of fluid-object interactions over long time periods while simultaneously making computation more efficient. Moreover, SurfsUP trained on simple shape primitives generalizes considerably out-of-distribution, even to complex real-world scenes and objects. Finally, we show we can invert our model to design simple objects to manipulate fluid flow.
****
Convolutional Networks with Oriented 1D Kernels
Alexandre Kirchmeyer, Jia Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6222-6232
In computer vision, 2D convolution is arguably the most important operation performed by a ConvNet. Unsurprisingly, it has been the focus of intense software and hardware optimization and enjoys highly efficient implementations.
In this work, we ask an intriguing question: can we make a ConvNet work without 2D convolutions? Surprisingly, we find that the answer is yes --- we show that a ConvNet consisting entirely of 1D convolutions can do just as well as 2D on ImageNet classification. Specifically, we find that one key ingredient to a high-performing 1D ConvNet is oriented 1D kernels: 1D kernels that are oriented not just horizontally or vertically, but also at other angles.
Our experiments show that oriented 1D convolutions can not only replace 2D convolutions but also augment existing architectures with large kernels, leading to improved accuracy with minimal FLOPs increase.
A key contribution of this work is a highly-optimized custom CUDA implementation of oriented 1D kernels, specialized to the depthwise convolution setting. Our benchmarks demonstrate that our custom CUDA implementation almost perfectly realizes the theoretical advantage of 1D convolution: it is faster than a native horizontal convolution for any arbitrary angle. Code is available at https://github.com/princeton-vl/Oriented1D.
****
Regularized Mask Tuning: Uncovering Hidden Knowledge in Pre-Trained Vision-Language Models
Kecheng Zheng, Wei Wu, Ruili Feng, Kai Zhu, Jiawei Liu, Deli Zhao, Zheng-Jun Zha, Wei Chen, Yujun Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11663-11673
Prompt tuning and adapter tuning have shown great potential in transferring pre-trained vision-language models (VLMs) to various downstream tasks. In this work, we design a new type of tuning method, termed as regularized mask tuning, which masks the network parameters through a learnable selection. Inspired by neural pathways, we argue that the knowledge required by a downstream task already exists in the pre-trained weights but just gets concealed in the upstream pre-training stage. To bring the useful knowledge back into light, we first identify a set of parameters that are important to a given downstream task, then attach a binary mask to each parameter, and finally optimize these masks on the downstream data with the parameters frozen. When updating the mask, we introduce a novel gradient dropout strategy to regularize the parameter selection, in order to prevent the model from forgetting old knowledge and overfitting the downstream data. Experimental results on 11 datasets demonstrate the consistent superiority of our method over previous alternatives. It is noteworthy that we manage to deliver 18.73% performance improvement compared to the zero-shot CLIP via masking an average of only 2.56% parameters. Furthermore, our method is synergistic with most existing parameter-efficient tuning methods and can boost the performance on top of them. Code will be made publicly available.

****
Skill Transformer: A Monolithic Policy for Mobile Manipulation
Xiaoyu Huang, Dhruv Batra, Akshara Rai, Andrew Szot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10852-10862
We present Skill Transformer, an approach for solving long-horizon robotic tasks by combining conditional sequence modeling and skill modularity. Conditioned on egocentric and proprioceptive observations of a robot, Skill Transformer is trained end-to-end to predict both a high-level skill (e.g., navigation, picking, placing), and a whole-body low-level action (e.g., base and arm motion), using a transformer architecture and demonstration trajectories that solve the full task. It retains the composability and modularity of the overall task through a skill predictor module while reasoning about low-level actions and avoiding hand-off errors, common in modular approaches. We test Skill Transformer on an embodied rearrangement benchmark and find it performs robust task planning and low-level control in new scenarios, achieving a 2.5x higher success rate than baselines in hard rearrangement problems.
****
Adaptive and Background-Aware Vision Transformer for Real-Time UAV Tracking
Shuiwang Li, Yangxiang Yang, Dan Zeng, Xucheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13989-14000
While discriminative correlation filters (DCF)-based trackers prevail in UAV tracking for their favorable efficiency, lightweight convolutional neural network (CNN)-based trackers using filter pruning have also demonstrated remarkable efficiency and precision. However, the use of pure vision transformer models (ViTs) for UAV tracking remains unexplored, which is a surprising finding given that ViTs have been shown to produce better performance and greater efficiency than CNNs in image classification. In this paper, we propose an efficient ViT-based tracking framework, Aba-ViTrack, for UAV tracking. In our framework, feature learning and template-search coupling are integrated into an efficient one-stream ViT to avoid an extra heavy relation modeling module. The proposed Aba-ViT exploits an adaptive and background-aware token computation method to reduce inference time. This approach adaptively discards tokens based on learned halting probabilities, which a priori are higher for background tokens than target ones. Extensive experiments on six UAV tracking benchmarks demonstrate that the proposed Aba-ViTrack achieves state-of-the-art performance in UAV tracking. Code is available at https://github.com/xyyang317/Aba-ViTrack.
****
Improving Pixel-based MIM by Reducing Wasted Modeling Capability
Yuan Liu, Songyang Zhang, Jiacheng Chen, Zhaohui Yu, Kai Chen, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5361-5372
There has been significant progress in Masked Image Modeling (MIM). Existing MIM methods can be broadly categorized into two groups based on the reconstruction target: pixel-based and tokenizer-based approaches. The former offers a simpler pipeline and lower computational cost, but it is known to be biased toward high-frequency details. In this paper, we provide a set of empirical studies to confirm this limitation of pixel-based MIM and propose a new method that explicitly utilizes low-level features from shallow layers to aid pixel reconstruction. By incorporating this design into our base method, MAE, we reduce the wasted modeling capability of pixel-based MIM, improving its convergence and achieving non-trivial improvements across various downstream tasks. To the best of our knowledge, we are the first to systematically investigate multi-level feature fusion for isotropic architectures like the standard Vision Transformer (ViT). Notably, when applied to a smaller model (e.g., ViT-S), our method yields significant performance gains, such as 1.2% on fine-tuning, 2.8% on linear probing, and 2.6% on semantic segmentation.
****
Towards Memory- and Time-Efficient Backpropagation for Training Spiking Neural Networks
Qingyan Meng, Mingqing Xiao, Shen Yan, Yisen Wang, Zhouchen Lin, Zhi-Quan Luo; P

roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6166-6176

Spiking Neural Networks (SNNs) are promising energy-efficient models for neuromorphic computing. For training the non-differentiable SNN models, the backpropagation through time (BPTT) with surrogate gradients (SG) method has achieved high performance. However, this method suffers from considerable memory cost and training time during training. In this paper, we propose the Spatial Learning Through Time (SLTT) method that can achieve high performance while greatly improving training efficiency compared with BPTT. First, we show that the backpropagation of SNNs through the temporal domain contributes just a little to the final calculated gradients. Thus, we propose to ignore the unimportant routes in the computational graph during backpropagation. The proposed method reduces the number of scalar multiplications and achieves a small memory occupation that is independent of the total time steps. Furthermore, we propose a variant of SLTT, called SLTT-K, that allows backpropagation only at K time steps, then the required number of scalar multiplications is further reduced and is independent of the total time steps. Experiments on both static and neuromorphic datasets demonstrate superior training efficiency and performance of our SLTT. In particular, our method achieves state-of-the-art accuracy on ImageNet, while the memory cost and training time are reduced by more than 70% and 50%, respectively, compared with BPTT.

****

Persistent-Transient Duality: A Multi-Mechanism Approach for Modeling Human-Object Interaction

Hung Tran, Vuong Le, Svetha Venkatesh, Truyen Tran; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9858-9867

Humans are highly adaptable, swiftly switching between different modes to progressively handle different tasks, situations and contexts. In Human-object interaction (HOI) activities, these modes can be attributed to two mechanisms: (1) the large-scale consistent plan for the whole activity and (2) the small-scale children interactive actions that start and end along the timeline. While neuroscience and cognitive science have confirmed this multi-mechanism nature of human behavior, machine modeling approaches for human motion are trailing behind. While attempted to use gradually morphing structures (e.g., graph attention networks) to model the dynamic HOI patterns, they miss the expeditious and discrete mode-switching nature of the human motion. To bridge that gap, this work proposes to model two concurrent mechanisms that jointly control human motion: the Persistent process that runs continually on the global scale, and the Transient sub-processes that operate intermittently on the local context of the human while interacting with objects. These two mechanisms form an interactive Persistent-Transient Duality that synergistically governs the activity sequences. We model this conceptual duality by a parent-child neural network of Persistent and Transient channels with a dedicated neural module for dynamic mechanism switching. The framework is trialed on HOI motion forecasting. On two rich datasets and a wide variety of settings, the model consistently delivers superior performances, proving its suitability for the challenge.

****

When to Learn What: Model-Adaptive Data Augmentation Curriculum

Chengkai Hou, Jieyu Zhang, Tianyi Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1717-1728

Data augmentation (DA) is widely used to improve the generalization of neural networks by enforcing the invariances and symmetries to pre-defined transformations applied to input data. However, a fixed augmentation policy may have different effects on each sample in different training stages but existing approaches cannot adjust the policy to be adaptive to each sample and the training model. In this paper, we propose "Model-Adaptive Data Augmentation (MADAug)" that jointly trains an augmentation policy network to teach the model "when to learn what". Unlike previous work, MADAug selects augmentation operators for each input image by a model-adaptive policy varying between training stages, producing a data augmentation curriculum optimized for better generalization. In MADAug, we train the policy through a bi-level optimization scheme, which aims to minimize a validat

ion set loss of a model trained using the policy-produced data augmentations. We conduct an extensive evaluation of MADAug on multiple image classification tasks and network architectures with thorough comparisons to existing DA approaches. MADAug outperforms or is on par with other baselines and exhibits better fairness: it brings improvement to all classes and more to the difficult ones. Moreover, MADAug learned policy shows better performance when transferred to fine-grained datasets. In addition, the auto-optimized policy in MADAug gradually introduces increasing perturbations and naturally forms an easy-to-hard curriculum.
****
DiffPose: Multi-hypothesis Human Pose Estimation using Diffusion Models
Karl Holmquist, Bastian Wandt; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15977-15987
Traditionally, monocular 3D human pose estimation employs a machine learning model to predict the most likely 3D pose for a given input image. However, a single image can be highly ambiguous and induces multiple plausible solutions for the 2D-3D lifting step, which results in overly confident 3D pose predictors. To this end, we propose DiffPose, a conditional diffusion model that predicts multiple hypotheses for a given input image. Compared to similar approaches, our diffusion model is straightforward and avoids intensive hyperparameter tuning, complex network structures, mode collapse, and unstable training.
Moreover, we tackle the problem of over-simplification of the intermediate representation of the common two-step approaches which first estimate a distribution of 2D joint locations via joint-wise heatmaps and consecutively use their maximum argument for the 3D pose estimation step. Since such a simplification of the heatmaps removes valid information about possibly correct, though labeled unlikely, joint locations, we propose to represent the heatmaps as a set of 2D joint candidate samples. To extract information about the original distribution from these samples, we introduce our embedding transformer which conditions the diffusion model. Experimentally, we show that DiffPose improves upon the state of the art for multi-hypothesis pose estimation by 3-5% for simple poses and outperforms it by a large
margin for highly ambiguous poses.
****
AesPA-Net: Aesthetic Pattern-Aware Style Transfer Networks
Kibeom Hong, Seogkyu Jeon, Junsoo Lee, Namhyuk Ahn, Kunhee Kim, Pilhyeon Lee, Daesik Kim, Youngjung Uh, Hyeran Byun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22758-22767
To deliver the artistic expression of the target style, recent studies exploit the attention mechanism owing to its ability to map the local patches of the style image to the corresponding patches of the content image. However, because of the low semantic correspondence between arbitrary content and artworks, the attention module repeatedly abuses specific local patches from the style image, resulting in disharmonious and evident repetitive artifacts. To overcome this limitation and accomplish impeccable artistic style transfer, we focus on enhancing the attention mechanism and capturing the rhythm of patterns that organize the style. In this paper, we introduce a novel metric, namely pattern repeatability, that quantifies the repetition of patterns in the style image. Based on the pattern repeatability, we propose Aesthetic Pattern-Aware style transfer Networks (AesPA-Net) that discover the sweet spot of local and global style expressions. In addition, we propose a novel self-supervisory task to encourage the attention mechanism to learn precise and meaningful semantic correspondence. Lastly, we introduce the patch-wise style loss to transfer the elaborate rhythm of local patterns. Through qualitative and quantitative evaluations, we verify the reliability of the proposed pattern repeatability that aligns with human perception, and demonstrate the superiority of the proposed framework.
****
COPILOT: Human-Environment Collision Prediction and Localization from Egocentric Videos
Boxiao Pan, Bokui Shen, Davis Rempe, Despoina Paschalidou, Kaichun Mo, Yanchao Yang, Leonidas J. Guibas; Proceedings of the IEEE/CVF International Conference on

Computer Vision (ICCV), 2023, pp. 5262-5272
The ability to forecast human-environment collisions from egocentric observations is vital to enable collision avoidance in applications such as VR, AR, and wearable assistive robotics. In this work, we introduce the challenging problem of predicting collisions in diverse environments from multi-view egocentric videos captured from body-mounted cameras. Solving this problem requires a generalizable perception system that can classify which human body joints will collide and estimate a collision region heatmap to localize collisions in the environment. To achieve this, we propose a transformer-based model called COPILOT to perform collision prediction and localization simultaneously, which accumulates information across multi-view inputs through a novel 4D space-time-viewpoint attention mechanism. To train our model and enable future research on this task, we develop a synthetic data generation framework that produces egocentric videos of virtual humans moving and colliding within diverse 3D environments. This framework is then used to establish a large-scale dataset consisting of 8.6M egocentric RGBD frames. Extensive experiments show that COPILOT generalizes to unseen synthetic as well as real-world scenes. We further demonstrate COPILOT outputs are useful for downstream collision avoidance through simple closed-loop control. Please visit our project webpage at https://sites.google.com/stanford.edu/copilot.
****
EGformer: Equirectangular Geometry-biased Transformer for 360 Depth Estimation
Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, Chae Eun Rhee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6101-6112
Estimating the depths of equirectangular (i.e., 360) images (EIs) is challenging given the distorted 180 x 360 field-of-view, which is hard to be addressed via convolutional neural network (CNN). Although a transformer with global attention achieves significant improvements over CNN for EI depth estimation task, it is computationally inefficient, which raises the need for transformer with local attention. However, to apply local attention successfully for EIs, a specific strategy, which addresses distorted equirectangular geometry and limited receptive field simultaneously, is required. Prior works have only cared either of them, resulting in unsatisfactory depths occasionally. In this paper, we propose an equirectangular geometry-biased transformer termed EGformer. While limiting the computational cost and the number of network parameters, EGformer enables the extraction of the equirectangular geometry-aware local attention with a large receptive field. To achieve this, we actively utilize the equirectangular geometry as the bias for the local attention instead of struggling to reduce the distortion of EIs. As compared to the most recent EI depth estimation studies, the proposed approach yields the best depth outcomes overall with the lowest computational cost and the fewest parameters, demonstrating the effectiveness of the proposed methods.
****
Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network
Chieh-Yun Chen, Yi-Chung Chen, Hong-Han Shuai, Wen-Huang Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7513-7522
Virtual try-on tasks aim at synthesizing realistic try-on results by trying target clothes on humans. Most previous works relied on the Thin Plate Spline or appearance flows to warp clothes to fit human body shapes. However, both approaches cannot handle complex warping, leading to over distortion or misalignment. Furthermore, there is a critical unaddressed challenge of adjusting clothing sizes for try-on. To tackle these issues, we propose a Clothing-Oriented Transformation Try-On Network (COTTON). COTTON leverages clothing structure with landmarks and segmentation to design a novel landmark-guided transformation for precisely deforming clothes, allowing for size adjustment during try-on. Additionally, to properly remove the clothing region from the human image without losing significant human characteristics, we propose a clothing elimination policy based on both transformed clothes and human segmentation. This method enables users to try on c

lothes tucked-in or untucked while retaining more human characteristics. Both qualitative and quantitative results show that COTTON outperforms the state-of-the-art high-resolution virtual try-on approaches. All the code is available at https://github.com/cotton6/COTTON-size-does-matter.
****
Generating Realistic Images from In-the-wild Sounds
Taegyeong Lee, Jeonghun Kang, Hyeonyu Kim, Taehwan Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7160-7170
Representing wild sounds as images is an important but challenging task due to the lack of paired datasets between sound and image data and the significant differences in the characteristics of these two modalities. Previous studies have focused on generating images from sound in limited categories or music. In this paper, we propose a novel approach to generate images from wild sounds. First, we convert sound into text using audio captioning. Second, we propose audio attention and sentence attention to represent the rich characteristics of sound and visualize the sound. Lastly, we propose a direct sound optimization with CLIPscore and AudioCLIP and generate images with a diffusion-based model. In experiments, it shows that our model is able to generate high quality images from wild sounds and outperforms baselines in both quantitative and qualitative evaluations on wild audio datasets.
****
DDS2M: Self-Supervised Denoising Diffusion Spatio-Spectral Model for Hyperspectral Image Restoration
Yuchun Miao, Lefei Zhang, Liangpei Zhang, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12086-12096
Diffusion models have recently received a surge of interest due to their impressive performance for image restoration, especially in terms of noise robustness. However, existing diffusion-based methods are trained on a large amount of training data and perform very well in-distribution, but can be quite susceptible to distribution shift. This is especially inappropriate for data-starved hyperspectral image (HSI) restoration. To tackle this problem, this work puts forth a self-supervised diffusion model for HSI restoration, namely Denoising Diffusion Spatio-Spectral Model (DDS2M), which works by inferring the parameters of the proposed Variational Spatio-Spectral Module (VS2M) during the reverse diffusion process, solely using the degraded HSI without any extra training data. In VS2M, a variational inference-based loss function is customized to enable the untrained spatial and spectral networks to learn the posterior distribution, which serves as the transitions of the sampling chain to help reverse the diffusion process. Benefiting from its self-supervised nature and the diffusion process, DDS2M enjoys stronger generalization ability to various HSIs compared to existing diffusion-based methods and superior robustness to noise compared to existing HSI restoration methods. Extensive experiments on HSI denoising, noisy HSI completion and super-resolution on a variety of HSIs demonstrate DDS2M's superiority over the existing task-specific state-of-the-arts. Code is available at: https://github.com/miaoyuchun/DDS2M.
****
Candidate-aware Selective Disambiguation Based On Normalized Entropy for Instance-dependent Partial-label Learning
Shuo He, Guowu Yang, Lei Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1792-1801
In partial-label learning (PLL), each training example has a set of candidate labels, among which only one is the true label. Most existing PLL studies focus on the instance-independent (II) case, where the generation of candidate labels is only dependent on the true label. However, this II-PLL paradigm could be unrealistic, since candidate labels are usually generated according to the specific features of the instance. Therefore, instance-dependent PLL (ID-PLL) has attracted increasing attention recently. Unfortunately, existing ID-PLL studies lack an insightful perception of the intrinsic challenge in ID-PLL. In this paper, we start with an empirical study of the dynamics of label disambiguation in both II-PLL and ID-PLL. We found that the performance degradation of ID-PLL stems from the

inaccurate supervision caused by massive under-disambiguated (UD) examples that do not achieve complete disambiguation. To solve this problem, we propose a novel two-stage PLL framework including selective disambiguation and candidate-aware thresholding. Specifically, we first choose a part of well-disambiguated (WD) examples based on the magnitude of normalized entropy (NE) and integrate harmless complementary supervision from the remaining ones to train two networks. Next, the remaining examples whose NE is lower than the specific class-wise WD-NE threshold are selected as additional WD ones. Meanwhile, the remaining UD examples, whose NE is lower than the self-adaptive UD-NE threshold and whose predictions from two networks are agreed, are also regarded as WD ones for model training. Extensive experiments demonstrate that our proposed method outperforms state-of-the-art PLL methods.
****
Open-vocabulary Video Question Answering: A New Benchmark for Evaluating the Generalizability of Video Question Answering Models
Dohwan Ko, Ji Soo Lee, Miso Choi, Jaewon Chu, Jihwan Park, Hyunwoo J. Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3101-3112
Video Question Answering (VideoQA) is a challenging task that entails complex multi-modal reasoning. In contrast to multiple-choice VideoQA which aims to predict the answer given several options, the goal of open-ended VideoQA is to answer questions without restricting candidate answers. However, the majority of previous VideoQA models formulate open-ended VideoQA as a classification task to classify the video-question pairs into a fixed answer set, i.e., closed-vocabulary, which contains only frequent answers (e.g., top-1000 answers). This leads the model to be biased toward only frequent answers and fail to generalize on out-of-vocabulary answers. We hence propose a new benchmark, Open-vocabulary Video Question Answering (OVQA), to measure the generalizability of VideoQA models by considering rare and unseen answers. In addition, in order to improve the model's generalization power, we introduce a novel GNN-based soft verbalizer that enhances the prediction on rare and unseen answers by aggregating the information from their similar words. For evaluation, we introduce new baselines by modifying the existing (closed-vocabulary) open-ended VideoQA models and improve their performances by further taking into account rare and unseen answers. Our ablation studies and qualitative analyses demonstrate that our GNN-based soft verbalizer further improves the model performance, especially on rare and unseen answers. We hope that our benchmark OVQA can serve as a guide for evaluating the generalizability of VideoQA models and inspire future research. Code is available at https://github.com/mlvlab/OVQA.
****
Using a Waffle Iron for Automotive Point Cloud Semantic Segmentation
Gilles Puy, Alexandre Boulch, Renaud Marlet; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3379-3389
Semantic segmentation of point clouds in autonomous driving datasets requires techniques that can process large numbers of points efficiently. Sparse 3D convolutions have become the de-facto tools to construct deep neural networks for this task: they exploit point cloud sparsity to reduce the memory and computational loads and are at the core of today's best methods. In this paper, we propose an alternative method that reaches the level of state-of-the-art methods without requiring sparse convolutions. We actually show that such level of performance is achievable by relying on tools a priori unfit for large scale and high-performing 3D perception. In particular, we propose a novel 3D backbone, WaffleIron, made almost exclusively of MLPs and dense 2D convolutions and present how to train it to reach high performance on SemanticKITTI and nuScenes. We believe that WaffleIron is a compelling alternative to backbones using sparse 3D convolutions, especially in frameworks and on hardware where those convolutions are not readily available.
****
AutoReP: Automatic ReLU Replacement for Fast Private Network Inference
Hongwu Peng, Shaoyi Huang, Tong Zhou, Yukui Luo, Chenghong Wang, Zigeng Wang, Ji

ahui Zhao, Xi Xie, Ang Li, Tony Geng, Kaleel Mahmood, Wujie Wen, Xiaolin Xu, Caiwen Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5178-5188

The growth of the Machine-Learning-As-A-Service (MLaaS) market has highlighted clients' data privacy and security issues. Private inference (PI) techniques using cryptographic primitives offer a solution but often have high computation and communication costs, particularly with non-linear operators like ReLU. Many attempts to reduce ReLU operations exist, but they may need heuristic threshold selection or cause substantial accuracy loss. This work introduces AutoReP, a gradient-based approach to lessen non-linear operators and alleviate these issues. It automates the selection of ReLU and polynomial functions to speed up PI applications and introduces distribution-aware polynomial approximation (DaPa) to maintain model expressivity while accurately approximating ReLUs. Our experimental results demonstrate significant accuracy improvements of 6.12% (94.31%, 12.9K ReLU budget, CIFAR-10), 8.39% (74.92%, 12.9K ReLU budget, CIFAR-100), and 9.45% (63.69%, 55K ReLU budget, Tiny-ImageNet) over current state-of-the-art methods, e.g., SNL. Morever, AutoReP is applied to EfficientNet-B2 on ImageNet dataset, and achieved 75.55% accuracy with 176.1 xReLU budget reduction.

****

MotionLM: Multi-Agent Motion Forecasting as Language Modeling

Ari Seff, Brian Cera, Dian Chen, Mason Ng, Aurick Zhou, Nigamaa Nayakanti, Khaled S. Refaat, Rami Al-Rfou, Benjamin Sapp; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8579-8590

Reliable forecasting of the future behavior of road agents is a critical component to safe planning in autonomous vehicles. Here, we represent continuous trajectories as sequences of discrete motion tokens and cast multi-agent motion prediction as a language modeling task over this domain. Our model, MotionLM, provides several advantages: First, it does not require anchors or explicit latent variable optimization to learn multimodal distributions. Instead, we leverage a single standard language modeling objective, maximizing the average log probability over sequence tokens. Second, our approach bypasses post-hoc interaction heuristics where individual agent trajectory generation is conducted prior to interactive scoring. Instead, MotionLM produces joint distributions over interactive agent futures in a single autoregressive decoding process. In addition, the model's sequential factorization enables temporally causal conditional rollouts. The proposed approach establishes new state-of-the-art performance for multi-agent motion prediction on the Waymo Open Motion Dataset, ranking 1st on the interactive challenge leaderboard.

****

Black Box Few-Shot Adaptation for Vision-Language Models

Yassine Ouali, Adrian Bulat, Brais Matinez, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15534-15546

Vision-Language (V-L) models trained with contrastive learning to align the visual and language modalities have been shown to be strong few-shot learners. Soft prompt learning is the method of choice for few-shot downstream adaption aiming to bridge the modality gap caused by the distribution shift induced by the new domain. While parameter-efficient, prompt learning still requires access to the model weights and can be computationally infeasible for large models with billions of parameters. To address these shortcomings, in this work, we describe a black-box method for V-L few-shot adaptation that (a) operates on pre-computed image and text features and hence works without access to the model's weights, (b) it is orders of magnitude faster at training time, (c) it is amenable to both supervised and unsupervised training, and (d) it can be even used to align image and text features computed from uni-modal models. To achieve this, we propose Linear Feature Alignment (LFA), a simple linear approach for V-L re-alignment in the target domain. LFA is initialized from a closed-form solution to a least-squares problem and then it is iteratively updated by minimizing a re-ranking loss. Despite its simplicity, our approach can even surpass soft-prompt learning methods as shown by extensive experiments on 11 image and 2 video datasets.

****
Center-Based Decoupled Point-cloud Registration for 6D Object Pose Estimation
Haobo Jiang, Zheng Dang, Shuo Gu, Jin Xie, Mathieu Salzmann, Jian Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3427-3437
In this paper, we propose a novel center-based decoupled point cloud registration framework for robust 6D object pose estimation in real-world scenarios. Our method decouples the translation from the entire transformation by predicting the object center and estimating the rotation in a center-aware manner. This center offset-based translation estimation is correspondence-free, freeing us from the difficulty of constructing correspondences in challenging scenarios, thus improving robustness. To obtain reliable center predictions, we use a multi-view (bird's eye view and front view) object shape description of the source-point features, with both views jointly voting for the object center. Additionally, we propose an effective shape embedding module to augment the source features, largely completing the missing shape information due to partial scanning, thus facilitating the center prediction. With the center-aligned source and model point clouds, the rotation predictor utilizes feature similarity to establish putative correspondences for SVD-based rotation estimation. In particular, we introduce a center-aware hybrid feature descriptor with a normal correction technique to extract discriminative, part-aware features for high-quality correspondence construction. Our experiments show that our method outperforms the state-of-the-art methods by a large margin on real-world datasets such as TUD-L, LINEMOD, and Occluded-LINEMOD.
****
Self-Ordering Point Clouds
Pengwan Yang, Cees G. M. Snoek, Yuki M. Asano; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15813-15822
In this paper we address the task of finding representative subsets of points in a 3D point cloud by means of a point-wise ordering. Only a few works have tried to address this challenging vision problem, all with the help of hard to obtain point and cloud labels. Different from these works, we introduce the task of point-wise ordering in 3D point clouds through self-supervision, which we call self-ordering. We further contribute the first end-to-end trainable network that learns a point-wise ordering in a self-supervised fashion. It utilizes a novel differentiable point scoring-sorting strategy and it constructs an hierarchical contrastive scheme to obtain self-supervision signals. We extensively ablate the method and show its superior performance even compared to supervised ordering methods on multiple datasets and tasks including zero-shot ordering of point clouds from unseen categories.
****
Continual Segment: Towards a Single, Unified and Non-forgetting Continual Segmentation Model of 143 Whole-body Organs in CT Scans
Zhanghexuan Ji, Dazhou Guo, Puyang Wang, Ke Yan, Le Lu, Minfeng Xu, Qifeng Wang, Jia Ge, Mingchen Gao, Xianghua Ye, Dakai Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21140-21151
Deep learning empowers the mainstream medical image segmentation methods. Nevertheless, current deep segmentation approaches are not capable of efficiently and effectively adapting and updating the trained models when new segmentation classes are incrementally added. In the real clinical environment, it can be preferred that segmentation models could be dynamically extended to segment new organs/tumors without the (re-)access to previous training datasets due to obstacles of patient privacy and data storage. This process can be viewed as a continual semantic segmentation (CSS) problem, being understudied for multi-organ segmentation. In this work, we propose a new architectural CSS learning framework to learn a single deep segmentation model for segmenting a total of 143 whole-body organs. Using the encoder/decoder network structure, we demonstrate that a continually trained then frozen encoder coupled with incrementally-added decoders can extract sufficiently representative image features for new classes to be subsequently and validly segmented, while avoiding the catastrophic forgetting in CSS. To mai

ntain a single network model complexity, each decoder is progressively pruned using neural architecture search and teacher-student based knowledge distillation. Finally, we propose a body-part and anomaly-aware output merging module to combine organ predictions originating from different decoders and incorporate both healthy and pathological organs appearing in different datasets. Trained and validated on 3D CT scans of 2500+ patients from four datasets, our single network can segment a total of 143 whole-body organs with very high accuracy, closely reaching the upper bound performance level by training four separate segmentation models (i.e., one model per dataset/task).
****

Enhancing Modality-Agnostic Representations via Meta-Learning for Brain Tumor Segmentation

Aishik Konwer, Xiaoling Hu, Joseph Bae, Xuan Xu, Chao Chen, Prateek Prasanna; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21415-21425

In medical vision, different imaging modalities provide complementary information. However, in practice, not all modalities may be available during inference or even training. Previous approaches, e.g., knowledge distillation or image synthesis, often assume the availability of full modalities for all patients during training; this is unrealistic and impractical due to the variability in data collection across sites. We propose a novel approach to learn enhanced modality-agnostic representations by employing a meta-learning strategy in training, even when only limited full modality samples are available. Meta-learning enhances partial modality representations to full modality representations by meta-training on partial modality data and meta-testing on limited full modality samples. Additionally, we co-supervise this feature enrichment by introducing an auxiliary adversarial learning branch. More specifically, a missing modality detector is used as a discriminator to mimic the full modality setting. Our segmentation framework significantly outperforms state-of-the-art brain tumor segmentation techniques in missing modality scenarios.
****

Zero-1-to-3: Zero-shot One Image to 3D Object

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, Carl Vondrick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9298-9309

We introduce Zero-1-to-3, a framework for changing the camera viewpoint of an object given just a single RGB image. To perform novel view synthesis in this underconstrained setting, we capitalize on the geometric priors that large-scale diffusion models learn about natural images. Our conditional diffusion model uses a synthetic dataset to learn controls of the relative camera viewpoint, which allow new images to be generated of the same object under a specified camera transformation. Even though it is trained on a synthetic dataset, our model retains a strong zero-shot generalization ability to out-of-distribution datasets as well as in-the-wild images, including impressionist paintings. Our viewpoint-conditioned diffusion approach can further be used for the task of 3D reconstruction from a single image. Qualitative and quantitative experiments show that our method significantly outperforms stateof- the-art single-view 3D reconstruction and novel view synthesis models by leveraging Internet-scale pre-training.
****

3D Distillation: Improving Self-Supervised Monocular Depth Estimation on Reflective Surfaces

Xuepeng Shi, Georgi Dikov, Gerhard Reitmayr, Tae-Kyun Kim, Mohsen Ghafoorian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9133-9143

Self-supervised monocular depth estimation (SSMDE) aims at predicting the dense depth maps of monocular images, by learning to minimize a photometric loss using spatially neighboring image pairs during training. While SSMDE offers a significant scalability advantage over supervised approaches, it performs poorly on reflective surfaces as the photometric constancy assumption of the photometric loss is violated. We note that the appearance of reflective surfaces is view-depende

nt and often there are views of such surfaces in the training data that are not contaminated by strong specular reflections. Thus, reflective surfaces can be accurately reconstructed by aggregating the predicted depth of these views. Motivated by this observation, we propose 3D distillation: a novel training framework that utilizes the projected depth of reconstructed reflective surfaces to generate reasonably accurate depth pseudo-labels. To identify those surfaces automatically, we employ an uncertainty-guided depth fusion method, combining the smoother and more accurate projected depth on reflective surfaces and the detailed predicted depth elsewhere. In our experiments using the ScanNet and 7-Scenes datasets, we show that 3D distillation not only significantly improves the prediction accuracy, especially on the problematic surfaces, but also that it generalizes well over various underlying network architectures and to new datasets.
****

GAIT: Generating Aesthetic Indoor Tours with Deep Reinforcement Learning
Desai Xie, Ping Hu, Xin Sun, Soren Pirk, Jianming Zhang, Radomir Mech, Arie E. Kaufman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7409-7419
Placing and orienting a camera to compose aesthetically meaningful shots of a scene is not only a key objective in real-world photography and cinematography but also for virtual content creation. The framing of a camera often significantly contributes to the story telling in movies, games, and mixed reality applications. Generating single camera poses or even contiguous trajectories either requires a significant amount of manual labor or requires solving high-dimensional optimization problems, which can be computationally demanding and error-prone. In this paper, we introduce GAIT, a framework for training a Deep Reinforcement Learning (DRL) agent, that learns to automatically control a camera to generate a sequence of aesthetically meaningful views for synthetic 3D indoor scenes. To generate sequences of frames with high aesthetic value, GAIT relies on a neural aesthetics estimator, which is trained on a crowed-sourced dataset. Additionally, we introduce regularization techniques for diversity and smoothness to generate visually interesting trajectories for a 3D environment, and to constrain agent acceleration in the reward function to generate a smooth sequence of camera frames. We validated our method by comparing it to baseline algorithms, based on a perceptual user study, and through ablation studies. Code and visual results are available on the project website: https://desaixie.github.io/gait-rl
****

Low-Light Image Enhancement with Multi-Stage Residue Quantization and Brightness-Aware Attention
Yunlong Liu, Tao Huang, Weisheng Dong, Fangfang Wu, Xin Li, Guangming Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12140-12149
Low-light image enhancement (LLIE) aims to recover illumination and improve the visibility of low-light images. Conventional LLIE methods often produce poor results because they neglect the effect of noise interference. Deep learning-based LLIE methods focus on learning a mapping function between low-light images and normal-light images that outperforms conventional LLIE methods. However, most deep learning-based LLIE methods cannot yet fully exploit the guidance of auxiliary priors provided by normal-light images in the training dataset. In this paper, we propose a brightness-aware network with normal-light priors based on brightness-aware attention and residual quantized codebook. To achieve a more natural and realistic enhancement, we design a query module to obtain more reliable normal-light features and fuse them with lowlight features by a fusion branch. In addition, we propose a brightness-aware attention module to further retain the color consistency between the enhanced results and the normal-light images. Extensive experimental results on both real-captured and synthetic data show that our method outperforms existing state-of-the-art methods.
****

Hierarchically Decomposed Graph Convolutional Networks for Skeleton-Based Action Recognition
Jungho Lee, Minhyeok Lee, Dogyoon Lee, Sangyoun Lee; Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV), 2023, pp. 10444-10453
Graph convolutional networks (GCNs) are the most commonly used methods for skele ton-based action recognition and have achieved remarkable performance. Generatin g adjacency matrices with semantically meaningful edges is particularly importan t for this task, but extracting such edges is challenging problem. To solve this , we propose a hierarchically decomposed graph convolutional network (HD-GCN) ar chitecture with a novel hierarchically decomposed graph (HD-Graph). The proposed HD-GCN effectively decomposes every joint node into several sets to extract maj or structurally adjacent and distant edges, and uses them to construct an HD-Gra ph containing those edges in the same semantic spaces of a human skeleton. In ad dition, we introduce an attention-guided hierarchy aggregation (A-HA) module to highlight the dominant hierarchical edge sets of the HD-Graph. Furthermore, we a pply a new six-way ensemble method, which uses only joint and bone stream withou t any motion stream. The proposed model is evaluated and achieves state-of-the-a rt performance on four large, popular datasets. Finally, we demonstrate the effe ctiveness of our model with various comparative experiments.
****
LIST: Learning Implicitly from Spatial Transformers for Single-View 3D Reconstru ction
Mohammad Samiul Arshad, William J. Beksi; Proceedings of the IEEE/CVF Internatio nal Conference on Computer Vision (ICCV), 2023, pp. 9321-9330
Accurate reconstruction of both the geometric and topological details of a 3D ob ject from a single 2D image embodies a fundamental challenge in computer vision. Existing explicit/implicit solutions to this problem struggle to recover self-o ccluded geometry and/or faithfully reconstruct topological shape structures. To resolve this dilemma, we introduce LIST, a novel neural architecture that levera ges local and global image features to accurately reconstruct the geometric and topological structure of a 3D object from a single image. We utilize global 2D f eatures to predict a coarse shape of the target object and then use it as a base for higher-resolution reconstruction. By leveraging both local 2D features from the image and 3D features from the coarse prediction, we can predict the signed distance between an arbitrary point and the target surface via an implicit pred ictor with great accuracy. Furthermore, our model does not require camera estima tion or pixel alignment. It provides an uninfluenced reconstruction from the inp ut-view direction. Through qualitative and quantitative analysis, we show the su periority of our model in reconstructing 3D objects from both synthetic and real -world images against the state of the art.
****
Rethinking Mobile Block for Efficient Attention-based Models
Jiangning Zhang, Xiangtai Li, Jian Li, Liang Liu, Zhucun Xue, Boshen Zhang, Zhen gkai Jiang, Tianxin Huang, Yabiao Wang, Chengjie Wang; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2023, pp. 1389-1400
This paper focuses on developing modern, efficient, lightweight models for dense predictions while trading off parameters, FLOPs, and performance. Inverted Resi dual Block (IRB) serves as the infrastructure for lightweight CNNs, but no count erpart has been recognized by attention-based studies. This work rethinks lightw eight infrastructure from efficient IRB and effective components of Transformer from a unified perspective, extending CNN-based IRB to attention-based models an d abstracting a one-residual Meta Mobile Block (MMB) for lightweight model desig n. Following simple but effective design criterion, we deduce a modern Inverted Residual Mobile Block (iRMB) and build a ResNet-like Efficient MOdel (EMO) with only iRMB for down-stream tasks. Extensive experiments on ImageNet-1K, COCO2017, and ADE20K benchmarks demonstrate the superiority of our EMO over state-of-the- art methods, e.g., EMO-1M/2M/5M achieve 71.5, 75.1, and 78.4 Top-1 that surpass equal-order CNN-/Attention-based models, while trading-off the parameter, effici ency, and accuracy well: running 2.8-4.0x faster than EdgeNeXt on iPhone14.
****
REAP: A Large-Scale Realistic Adversarial Patch Benchmark
Nabeel Hingun, Chawin Sitawarin, Jerry Li, David Wagner; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2023, pp. 4640-4651

Machine learning models are known to be susceptible to adversarial perturbation. One famous attack is the adversarial patch, a particularly crafted sticker that makes the model mispredict the object it is placed on. This attack presents a critical threat to cyber-physical systems that rely on cameras such as autonomous cars. Despite the significance of the problem, conducting research in this setting has been difficult; evaluating attacks and defenses in the real world is exceptionally costly while synthetic data are unrealistic. In this work, we propose the REAP (REalistic Adversarial Patch) benchmark, a digital benchmark that enables the evaluations on real images under real-world conditions. Built on top of the Mapillary Vistas dataset, our benchmark contains over 14,000 traffic signs. Each sign is augmented with geometric and lighting transformations for applying a digitally generated patch realistically onto the sign. Using our benchmark, we perform the first large-scale assessments of adversarial patch attacks under realistic conditions. Our experiments suggest that patch attacks may present a smaller threat than previously believed and that the success rate of an attack on simpler digital simulations is not predictive of its actual effectiveness in practice. Our benchmark is released publicly at https://github.com/wagner-group/reap-benchmark.
****

LRRU: Long-short Range Recurrent Updating Networks for Depth Completion
Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, Yuchao Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9422-9432
Existing deep learning-based depth completion methods generally employ massive stacked layers to predict the dense depth map from sparse input data. Although such approaches greatly advance this task, their accompanied huge computational complexity hinders their practical applications. To accomplish depth completion more efficiently, we propose a novel lightweight deep network framework, the Long-short Range Recurrent Updating (LRRU) network. Without learning complex feature representations, LRRU first roughly fills the sparse input to obtain an initial dense depth map, and then iteratively updates it through learned spatially-variant kernels. Our iterative update process is content-adaptive and highly flexible, where the kernel weights are learned by jointly considering the guidance RGB images and the depth map to be updated, and large-to-small kernel scopes are dynamically adjusted to capture long-to-short range dependencies. Our initial depth map has coarse but complete scene depth information, which helps relieve the burden of directly regressing the dense depth from sparse ones, while our proposed method can effectively refine it to an accurate depth map with less learnable parameters and inference time. Experimental results demonstrate that our proposed LRRU variants achieve state-of-the-art performance across different parameter regimes. In particular, the LRRU-Base model outperforms competing approaches on the NYUv2 dataset, and ranks 1st on the KITTI depth completion benchmark at the time of submission. Project page: https://npucvr.github.io/LRRU/.
****

MetaBEV: Solving Sensor Failures for 3D Detection and Map Segmentation
Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8721-8731
Perception systems in modern autonomous driving vehicles typically take inputs from complementary multi-modal sensors, e.g., LiDAR and cameras. However, in real-world applications, sensor corruptions and failures lead to inferior performances, thus compromising autonomous safety. In this paper, we propose a robust framework, called MetaBEV, to address extreme real-world environments, involving overall six sensor corruptions and two extreme sensor-missing situations. In MetaBEV, signals from multiple sensors are first processed by modal-specific encoders. Subsequently, a set of dense BEV queries are initialized, termed meta-BEV. These queries are then processed iteratively by a BEV-Evolving decoder, which selectively aggregates deep features from either LiDAR, cameras, or both modalities. The updated BEV representations are further leveraged for multiple 3D prediction tasks. Additionally, we introduce a new \moe   structure to alleviate the performance drop on distinct tasks in multi-task joint learning. Finally, MetaBEV is e

valuated on the nuScenes dataset with 3D object detection and BEV map segmentation tasks. Experiments show MetaBEV outperforms prior arts by a large margin on both full and corrupted modalities. For instance, when the LiDAR signal is missing, MetaBEV improves 35.5% detection NDS and 17.7% segmentation mIoU upon the vanilla BEVFusion model; and when the camera signal is absent, MetaBEV still achieves 69.2% NDS and 53.7%mIoU, which is even higher than previous works that perform on full-modalities. Moreover, MetaBEV performs moderately against previous methods in both canonical perception and multi-task learning settings, refreshing state-of-the-art nuScenes BEV map segmentation with 70.4% mIoU.
****
DNA-Rendering: A Diverse Neural Actor Repository for High-Fidelity Human-Centric Rendering
Wei Cheng, Ruixiang Chen, Siming Fan, Wanqi Yin, Keyu Chen, Zhongang Cai, Jingbo Wang, Yang Gao, Zhengming Yu, Zhengyu Lin, Daxuan Ren, Lei Yang, Ziwei Liu, Chen Change Loy, Chen Qian, Wayne Wu, Dahua Lin, Bo Dai, Kwan-Yee Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19982-19993
Realistic human-centric rendering plays a key role in both computer vision and computer graphics. Rapid progress has been made in the algorithm aspect over the years, yet existing human-centric rendering datasets and benchmarks are rather impoverished in terms of diversity (e.g., outfit's fabric/material, body's interaction with objects, and motion sequences), which are crucial for rendering effect. Researchers are usually constrained to explore and evaluate a small set of rendering problems on current datasets, while real-world applications require methods to be robust across different scenarios. In this work, we present DNA-Rendering, a large-scale, high-fidelity repository of human performance data for neural actor rendering. DNA-Rendering presents several appealing attributes. First, our dataset contains over 1500 human subjects, 5000 motion sequences, and 67.5M frames' data volume. Upon the massive collections, we provide human subjects with grand categories of pose actions, body shapes, clothing, accessories, hairdos, and object intersection, which ranges the geometry and appearance variances from everyday life to professional occasions. Second, we provide rich assets for each subject - 2D/3D human body keypoints, foreground masks, SMPLX models, cloth/accessory materials, multi-view images, and videos. These assets boost the current method's accuracy on downstream rendering tasks. Third, we construct a professional multi-view system to capture data, which contains 60 synchronous cameras with max 4096 x 3000 resolution, 15 fps speed, and stern camera calibration steps, ensuring high-quality resources for task training and evaluation.

Along with the dataset, we provide a large-scale and quantitative benchmark in full-scale, with multiple tasks to evaluate the existing progress of novel view synthesis, novel pose animation synthesis, and novel identity rendering methods. In this manuscript, we describe our DNA-Rendering effort as a revealing of new observations, challenges, and future directions to human-centric rendering. The dataset, code, and benchmarks will be publicly available at https://dna-rendering.github.io/.
****
Exploring Temporal Concurrency for Video-Language Representation Learning
Heng Zhang, Daqing Liu, Zezhong Lv, Bing Su, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15568-15578
Paired video and language data is naturally temporal concurrency, which requires the modeling of the temporal dynamics within each modality and the temporal alignment across modalities simultaneously. However, most existing video-language representation learning methods only focus on discrete semantic alignment that encourages aligned semantics to be close in the latent space, or temporal context dependency that captures short-range coherence, failing in building the temporal concurrency. In this paper, we propose to learn video-language representations by modeling video-language pairs as Temporal Concurrent Processes (TCP) via a process-wised distance metric learning framework. Specifically, we employ the soft Dynamic Time Warping (DTW) to measure the distance between two processes across

modalities and then optimize the DTW costs. Meanwhile, we further introduce a regularization term that enforces the embeddings of each modality approximating a stochastic process to guarantee the inherent dynamics. Experimental results on three benchmarks demonstrate that TCP stands as a state-of-the-art method for various video-language understanding tasks, including paragraph-to-video retrieval, video moment retrieval, and video question-answering. Code is available at https://github.com/hengRUC/TCP.
****
StegaNeRF: Embedding Invisible Information within Neural Radiance Fields
Chenxin Li, Brandon Y. Feng, Zhiwen Fan, Panwang Pan, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 441-453
Recent advancements in neural rendering have paved the way for a future marked by the widespread distribution of visual data through the sharing of Neural Radiance Field (NeRF) model weights. However, while established techniques exist for embedding ownership or copyright information within conventional visual data such as images and videos, the challenges posed by the emerging NeRF format have remained unaddressed. In this paper, we introduce StegaNeRF, an innovative approach for steganographic information embedding within NeRF renderings. We have meticulously developed an optimization framework that enables precise retrieval of hidden information from images generated by NeRF, while ensuring the original visual quality of the rendered images to remain intact. Through rigorous experimentation, we assess the efficacy of our methodology across various potential deployment scenarios. Furthermore, we delve into the insights gleaned from our analysis. StegaNeRF represents an initial foray into the intriguing realm of infusing NeRF renderings with customizable, imperceptible, and recoverable information, all while minimizing any discernible impact on the rendered images. For more details, please visit our project page: https://xggnet.github.io/StegaNeRF/
****
DynamicISP: Dynamically Controlled Image Signal Processor for Image Recognition
Masakazu Yoshimura, Junji Otsuka, Atsushi Irie, Takeshi Ohashi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12866-12876
Image Signal Processors (ISPs) play important roles in image recognition tasks as well as in the perceptual quality of captured images. In most cases, experts make a lot of effort to manually tune many parameters of ISPs, but the parameters are sub-optimal. In the literature, two types of techniques have been actively studied: a machine learning-based parameter tuning technique and a DNN-based ISP technique. The former is lightweight but lacks expressive power. The latter has expressive power, but the computational cost is too heavy on edge devices. To solve these problems, we propose "DynamicISP," which consists of multiple classical ISP functions and dynamically controls the parameters of each frame according to the recognition result of the previous frame. We show our method successfully controls the parameters of multiple ISP functions and achieves state-of-the-art accuracy with low computational cost in single and multi-category object detection tasks.
****
R-Pred: Two-Stage Motion Prediction Via Tube-Query Attention-Based Trajectory Refinement
Sehwan Choi, Jungho Kim, Junyong Yun, Jun Won Choi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8525-8535
Predicting the future motion of dynamic agents is of paramount importance to ensuring safety and assessing risks in motion planning for autonomous robots. In this study, we propose a two-stage motion prediction method, called R-Pred, designed to effectively utilize both scene and interaction context using a cascade of the initial trajectory proposal and trajectory refinement networks. The initial trajectory proposal network produces M trajectory proposals corresponding to the M modes of the future trajectory distribution. The trajectory refinement network enhances each of the M proposals using 1) tube-query scene attention (TQSA) and 2) proposal-level interaction attention (PIA) mechanisms. TQSA uses tube-queri

es to aggregate local scene context features pooled from proximity around trajectory proposals of interest. PIA further enhances the trajectory proposals by modeling inter-agent interactions using a group of trajectory proposals selected by their distances from neighboring agents. Our experiments conducted on Argoverse and nuScenes datasets demonstrate that the proposed refinement network provides significant performance improvements compared to the single-stage baseline and that R-Pred achieves state-of-the-art performance in some categories of the benchmarks.
****
A step towards understanding why classification helps regression
Silvia L. Pintea, Yancong Lin, Jouke Dijkstra, Jan C. van Gemert; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19972-19981
A number of computer vision deep regression approaches report improved results when adding a classification loss to the regression loss. Here, we explore why this is useful in practice and when it is beneficial. To do so, we start from precisely controlled dataset variations and data samplings and find that the effect of adding a classification loss is the most pronounced for regression with imbalanced data. We explain these empirical findings by formalizing the relation between the balanced and imbalanced regression losses. Finally, we show that our findings hold on two real imbalanced image datasets for depth estimation (NYUD2-DIR), and age estimation (IMDB-WIKI-DIR), and on the problem of imbalanced video progress prediction (Breakfast). Our main takeaway is: for a regression task, if the data sampling is imbalanced, then add a classification loss.
****
Robust Evaluation of Diffusion-Based Adversarial Purification
Minjong Lee, Dongwoo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 134-144
We question the current evaluation practice on diffusion-based purification methods. Diffusion-based purification methods aim to remove adversarial effects from an input data point at test time. The approach gains increasing attention as an alternative to adversarial training due to the disentangling between training and testing. Well-known white-box attacks are often employed to measure the robustness of the purification. However, it is unknown whether these attacks are the most effective for the diffusion-based purification since the attacks are often tailored for adversarial training. We analyze the current practices and provide a new guideline for measuring the robustness of purification methods against adversarial attacks. Based on our analysis, we further propose a new purification strategy improving robustness compared to the current diffusion-based purification methods.
****
Hyperbolic Audio-visual Zero-shot Learning
Jie Hong, Zeeshan Hayder, Junlin Han, Pengfei Fang, Mehrtash Harandi, Lars Petersson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7873-7883
Audio-visual zero-shot learning aims to classify samples consisting of a pair of corresponding audio and video sequences from classes that are not present during training. An analysis of the audio-visual data reveals a large degree of hyperbolicity, indicating the potential benefit of using a hyperbolic transformation to achieve curvature-aware geometric learning, with the aim of exploring more complex hierarchical data structures for this task. The proposed approach employs a novel loss function that incorporates cross-modality alignment between video and audio features in the hyperbolic space. Additionally, we explore the use of multiple adaptive curvatures for hyperbolic projections. The experimental results on this very challenging task demonstrate that our proposed hyperbolic approach for zero-shot learning outperforms the SOTA method on three datasets: VGGSound-GZSL, UCF-GZSL, and ActivityNet-GZSL achieving a harmonic mean (HM) improvement of around 3.0%, 7.0%, and 5.3%, respectively.
****
CTP:Towards Vision-Language Continual Pretraining via Compatible Momentum Contra

st and Topology Preservation
Hongguang Zhu, Yunchao Wei, Xiaodan Liang, Chunjie Zhang, Yao Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22257-22267
Vision-Language Pretraining (VLP) has shown impressive results on diverse downstream tasks by offline training on large-scale datasets. Regarding the growing nature of real-world data, such an offline training paradigm on ever-expanding data is unsustainable, because models lack the continual learning ability to accumulate knowledge constantly. However, most continual learning studies are limited to uni-modal classification and existing multi-modal datasets cannot simulate continual non-stationary data stream scenarios. To support the study of Vision-Language Continual Pretraining (VLCP), we first contribute a comprehensive and unified benchmark dataset P9D which contains over one million product image-text pairs from 9 industries. The data from each industry as an independent task supports continual learning and conforms to the real-world long-tail nature to simulate pretraining on web data. We comprehensively study the characteristics and challenges of VLCP, and propose a new algorithm: Compatible momentum contrast with Topology Preservation, dubbed CTP. The compatible momentum model absorbs the knowledge of the current and previous-task models to flexibly update the modal feature. Moreover, Topology Preservation transfers the knowledge of embedding across tasks while preserving the flexibility of feature adjustment. The experimental results demonstrate our method not only achieves superior performance compared with other baselines but also does not bring an expensive training burden. Dataset and codes are available at https://github.com/KevinLight831/CTP.
****
Aggregating Feature Point Cloud for Depth Completion
Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, Hui-Liang Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8732-8743
Guided depth completion aims to recover dense depth maps by propagating depth information from the given pixels to the remaining ones under the guidance of RGB images. However, most of the existing methods achieve this using a large number of iterative refinements or stacking repetitive blocks. Due to the limited receptive field of conventional convolution, the generalizability with respect to different sparsity levels of input depth maps is impeded. To tackle these problems, we propose a feature point cloud aggregation framework to directly propagate 3D depth information between the given points and the missing ones. We extract 2D feature map from images and transform the sparse depth map to point cloud to extract sparse 3D features. By regarding the extracted features as two sets of feature point clouds, the depth information for a target location can be reconstructed by aggregating adjacent sparse 3D features from the known points using cross attention. Based on this, we design a neural network, called as PointDC, to complete the entire depth information reconstruction process. Experimental results show that, our PointDC achieves superior or competitive results on the KITTI benchmark and NYUv2 dataset. In addition, the proposed PointDC demonstrates its higher generalizability to different sparsity levels of the input depth maps and cross-dataset evaluation.
****
FLIP: Cross-domain Face Anti-spoofing with Language Guidance
Koushik Srivatsan, Muzammal Naseer, Karthik Nandakumar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19685-19696
Face anti-spoofing (FAS) or presentation attack detection is an essential component of face recognition systems deployed in security-critical applications. Existing FAS methods have poor generalizability to unseen spoof types, camera sensors, and environmental conditions. Recently, vision transformer (ViT) models have been shown to be effective for the FAS task due to their ability to capture long-range dependencies among image patches. However, adaptive modules or auxiliary loss functions are often required to adapt pre-trained ViT weights learned on large-scale datasets such as ImageNet. In this work, we first show that initializing ViTs with multimodal (e.g., CLIP) pre-trained weights improves generalizabili

ty for the FAS task, which is in line with the zero-shot transfer capabilities of vision-language pre-trained (VLP) models. We then propose a novel approach for robust cross-domain FAS by grounding visual representations with the help of natural language. Specifically, we show that aligning the image representation with an ensemble of class descriptions (based on natural language semantics) improves FAS generalizability in low-data regimes. Finally, we propose a multimodal contrastive learning strategy to boost feature generalization further and bridge the gap between source and target domains. Extensive experiments on three standard protocols demonstrate that our method significantly outperforms the state-of-the-art methods, achieving better zero-shot transfer performance than five-shot transfer of "adaptive ViTs". Code: https://github.com/koushiksrivats/FLIP
****
Distribution Shift Matters for Knowledge Distillation with Webly Collected Images
Jialiang Tang, Shuo Chen, Gang Niu, Masashi Sugiyama, Chen Gong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17470-17480
Knowledge distillation aims to learn a lightweight student network from a pre-trained teacher network. In practice, existing knowledge distillation methods are usually infeasible when the original training data is unavailable due to some privacy issues and data management considerations. Therefore, data-free knowledge distillation approaches proposed to collect training instances from the Internet. However, most of them have ignored the common distribution shift between the instances from original training data and webly collected data, affecting the reliability of the trained student network. To solve this problem, we propose a novel method dubbed "Knowledge Distillation between Different Distributions" (KD^ 3), which consists of three components. Specifically, we first dynamically select useful training instances from the webly collected data according to the combined predictions of teacher network and student network. Subsequently, we align both the weighted features and classifier parameters of the two networks for knowledge memorization. Meanwhile, we also build a new contrastive learning block called MixDistribution to generate perturbed data with a new distribution for instance alignment, so that the student network can further learn a distribution-invariant representation. Intensive experiments on various benchmark datasets demonstrate that our proposed KD^ 3 can outperform the state-of-the-art data-free knowledge distillation approaches.
****
Reconstructed Convolution Module Based Look-Up Tables for Efficient Image Super-Resolution
Guandu Liu, Yukang Ding, Mading Li, Ming Sun, Xing Wen, Bin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12217-12226
Look-up table (LUT)-based methods have shown the great efficacy in single image super-resolution (SR) task.
However, previous methods don't delve into the essential reason of restricted receptive field (RF) size in LUT, which is caused by the interaction of space and channel features in vanilla convolution.
To enlarge RF with contained LUT sizes, we propose a novel Reconstructed Convolution(RC) module, which decouples channel-wise and spatial calculation. It can be formulated as n^2 1D LUTs to maintain nxn receptive field, which is obviously smaller than nxn D LUT formulated before. The LUT generated by our RC module reaches less than 1/10000 storage compared with SR-LUT baseline. The proposed Reconstructed Convolution module based LUT method, termed as RCLUT, can enlarge the RF size by 9 times than the state-of-the-art LUT-based SR method and achieve superior performance on five popular benchmark dataset. Moreover, the efficient and robust RC module can be used as a plugin to improve other LUT-based SR methods. The code is available at https://github.com/RC-LUT/RC-LUT.git.
****
Action Sensitivity Learning for Temporal Action Localization
Jiayi Shao, Xiaohan Wang, Ruijie Quan, Junjun Zheng, Jiang Yang, Yi Yang; Procee

dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13457-13469
Temporal action localization (TAL), which involves recognizing and locating action instances, is a challenging task in video understanding.
 Most existing approaches directly predict action classes and regress offsets to boundaries, while overlooking the discrepant importance of each frame.
 In this paper, we propose an Action Sensitivity Learning framework (ASL) to tackle this task, which aims to assess the value of each frame and then leverage the generated action sensitivity to recalibrate the training procedure. We first introduce a lightweight Action Sensitivity Evaluator to learn the action sensitivity at the class level and instance level, respectively. The outputs of the two branches are combined to reweight the gradient of the two sub-tasks. Moreover, based on the action sensitivity of each frame, we design an Action Sensitive Contrastive Loss to enhance features, where the action-aware frames are sampled as positive pairs to push away the action-irrelevant frames. The extensive studies on various action localization benchmarks (i.e., MultiThumos, Charades, Ego4D-Moment Queries v1.0, Epic-Kitchens 100, Thumos14 and ActivityNet1.3) show that ASL surpasses the state-of-the-art in terms of average-mAP under multiple types of scenarios, e.g., single-labeled, densely-labeled and egocentric.
****
Gram-based Attentive Neural Ordinary Differential Equations Network for Video Nystagmography Classification
Xihe Qiu, Shaojie Shi, Xiaoyu Tan, Chao Qu, Zhijun Fang, Hailing Wang, Yongbin Gao, Peixia Wu, Huawei Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21339-21348
Video nystagmography (VNG) is the diagnostic gold standard of benign paroxysmal positional vertigo (BPPV), which requires medical professionals to examine the direction, frequency, intensity, duration, and variation in the strength of nystagmus on a VNG video. This is a tedious process heavily influenced by the doctor's experience, which is error-prone. Recent automatic VNG classification methods approach this problem from the perspective of video analysis without considering medical prior knowledge, resulting in unsatisfactory accuracy and limited diagnostic capability for nystagmographic types, thereby preventing their clinical application. In this paper, we propose an end-to-end data-driven novel BPPV diagnosis framework (TC-BPPV) by considering this problem as an eye trajectory classification problem due to the disease's symptoms and experts' prior knowledge. In this framework, we utilize an eye movement tracking system to capture the eye trajectory and propose the Gram-based attentive neural ordinary differential equations network (Gram-AODE) to perform classification. We validate our framework using the VNG dataset provided by the collaborative university hospital and achieve state-of-the-art performance. We also evaluate Gram-AODE on multiple open-source benchmarks to demonstrate its effectiveness in trajectory classification. Code is available at https://github.com/XiheQiu/Gram-AODE.
****
PEANUT: Predicting and Navigating to Unseen Targets
Albert J. Zhai, Shenlong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10926-10935
Efficient ObjectGoal navigation (ObjectNav) in novel environments requires an understanding of the spatial and semantic regularities in environment layouts. In this work, we present a straightforward method for learning these regularities by predicting the locations of unobserved objects from incomplete semantic maps. Our method differs from previous prediction-based navigation methods, such as frontier potential prediction or egocentric map completion, by directly predicting unseen targets while leveraging the global context from all previously explored areas. Our prediction model is lightweight and can be trained in a supervised manner using a relatively small amount of passively collected data. Once trained, the model can be incorporated into a modular pipeline for ObjectNav without the need for any reinforcement learning. We validate the effectiveness of our method on the HM3D and MP3D ObjectNav datasets. We find that it achieves the state-of-the-art on both datasets, despite not using any additional data for training.

****
Pluralistic Aging Diffusion Autoencoder
Peipei Li, Rui Wang, Huaibo Huang, Ran He, Zhaofeng He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22613-22623
Face aging is an ill-posed problem because multiple plausible aging patterns may correspond to a given input. Most existing methods often produce one deterministic estimation. This paper proposes a novel CLIP-driven Pluralistic Aging Diffusion Autoencoder (PADA) to enhance the diversity of aging patterns. First, we employ diffusion models to generate diverse low-level aging details via a sequential denoising reverse process. Second, we present Probabilistic Aging Embedding (PAE) to capture diverse high-level aging patterns, which represents age information as probabilistic distributions in the common CLIP latent space. A text-guided KL-divergence loss is designed to guide this learning. Our method can achieve pluralistic face aging conditioned on open-world aging texts and arbitrary unseen face images. Qualitative and quantitative experiments demonstrate that our method can generate more diverse and high-quality plausible aging results.
****
ModelGiF: Gradient Fields for Model Functional Distance
Jie Song, Zhengqi Xu, Sai Wu, Gang Chen, Mingli Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6125-6135
The last decade has witnessed the success of deep learning and the surge of publicly released trained models, which necessitates the quantification of the model functional distance for various purposes. However, quantifying the model functional distance is always challenging due to the opacity in inner workings and the heterogeneity in architectures and tasks. Inspired by the concept of "field" in physics, in this work we introduce Model Gradient Field (abbr. ModelGiF) to extract homogeneous representations from the heterogeneous pre-trained models. Our main assumption underlying ModelGiF is that each pre-trained deep model uniquely determines a ModelGiF over the input space. The distance between models can thus be measured by the similarity between their ModelGiFs. We provide theoretical insights into the proposed ModelGiFs for model functional distance, and validate the effectiveness of the proposed ModelGiF with a suite of testbeds, including task relatedness estimation, intellectual property protection, and model unlearning verification. Experimental results demonstrate the versatility of the proposed ModelGiF on these tasks, with significantly superiority performance to state-of-the-art competitors. Codes are available at https://github.com/zju-vipa/modelgif.
****
PoseDiffusion: Solving Pose Estimation via Diffusion-aided Bundle Adjustment
Jianyuan Wang, Christian Rupprecht, David Novotny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9773-9783
Camera pose estimation is a long-standing computer vision problem that to date often relies on classical methods, such as handcrafted keypoint matching, RANSAC and bundle adjustment. In this paper, we propose to formulate the Structure from Motion (SfM) problem inside a probabilistic diffusion framework, modelling the conditional distribution of camera poses given input images. This novel view of an old problem has several advantages. (i) The nature of the diffusion framework mirrors the iterative procedure of bundle adjustment. (ii) The formulation allows a seamless integration of geometric constraints from epipolar geometry. (iii) It excels in typically difficult scenarios such as sparse views with wide baselines. (iv) The method can predict intrinsics and extrinsics for an arbitrary amount of images. We demonstrate that our method PoseDiffusion significantly improves over the classic SfM pipelines and the learned approaches on two real-world datasets. Finally, it is observed that our method can generalize across datasets without further training. Project page: https://posediffusion.github.io/
****
TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering
Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, Noah A. Smith; Proceedings of the IEEE/CVF International Conference on Computer

Vision (ICCV), 2023, pp. 20406-20417
Despite thousands of researchers, engineers, and artists actively working on improving text-to-image generation models, systems often fail to produce images that accurately align with the text inputs. We introduce TIFA (Text-to-image Faithfulness evaluation with question Answering), an automatic evaluation metric that measures the faithfulness of a generated image to its text input via visual question answering (VQA). Specifically, given a text input, we automatically generate several question-answer pairs using a language model. We calculate image faithfulness by checking whether existing VQA models can answer these questions using the generated image. TIFA is a reference-free metric that allows for fine-grained and interpretable evaluations of generated images.TIFA also has better correlations with human judgments than existing metrics. Based on this approach, we introduce TIFA v1.0, a benchmark consisting of 4K diverse text inputs and 25K questions across 12 categories (object, counting, etc.). We present a comprehensive evaluation of existing text-to-image models using TIFA v1.0 and highlight the limitations and challenges of current models. For instance, we find that current text-to-image models, despite doing well on color and material, still struggle in counting, spatial relations, and composing multiple objects. We hope our benchmark will help carefully measure the research progress in text-to-image synthesis and provide valuable insights for further research.
****
SIGMA: Scale-Invariant Global Sparse Shape Matching
Maolin Gao, Paul Roetzer, Marvin Eisenberger, Zorah Lähner, Michael Moeller, Daniel Cremers, Florian Bernard; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 645-654
We propose a novel mixed-integer programming (MIP) formulation for generating precise sparse correspondences for highly non-rigid shapes. To this end, we introduce a projected Laplace-Beltrami operator (PLBO) which combines intrinsic and extrinsic geometric information to measure the deformation quality induced by predicted correspondences. We integrate the PLBO, together with an orientation-aware regulariser, into a novel MIP formulation that can be solved to global optimality for many practical problems. In contrast to previous methods, our approach is provably invariant to rigid transformations and global scaling, initialisation-free, has optimality guarantees, and scales to high resolution meshes with (empirically observed) linear time. We show state-of-the-art results for sparse non-rigid matching on several challenging 3D datasets, including data with inconsistent meshing, as well as applications in mesh-to-point-cloud matching.
****
CORE: Cooperative Reconstruction for Multi-Agent Perception
Binglu Wang, Lei Zhang, Zhaozhong Wang, Yongqiang Zhao, Tianfei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8710-8720
This paper presents CORE, a conceptually simple, effective and communication-efficient model for multi-agent cooperative perception. It addresses the task from a novel perspective of cooperative reconstruction, based on two key insights: 1) cooperating agents together provide a more holistic observation of the environment, and 2) the holistic observation can serve as valuable supervision to explicitly guide the model learning how to reconstruct the ideal observation based on collaboration. CORE instantiates the idea with three major components: a compressor for each agent to create more compact feature representation for efficient broadcasting, a lightweight attentive collaboration component for cross-agent message aggregation, and a reconstruction module to reconstruct the observation based on aggregated feature representations. This learning-to-reconstruct idea is task-agnostic, and offers clear and reasonable supervision to inspire more effective collaboration, eventually promoting perception tasks. We validate CORE on two large-scale multi-agent percetion dataset, OPV2V and V2X-Sim, in two tasks, i.e., 3D object detection and semantic segmentation. Results demonstrate that CORE achieves state-of-the-art performance, and is more communication-efficient.
****
VidStyleODE: Disentangled Video Editing via StyleGAN and NeuralODEs

Moayed Haji Ali, Andrew Bond, Tolga Birdal, Duygu Ceylan, Levent Karacan, Erkut Erdem, Aykut Erdem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7523-7534
We propose VidStyleODE, a spatiotemporally continuous disentangled video representation based upon StyleGAN and Neural-ODEs. Effective traversal of the latent space learned by Generative Adversarial Networks (GANs) has been the basis for recent breakthroughs in image editing. However, the applicability of such advancements to the video domain has been hindered by the difficulty of representing and controlling videos in the latent space of GANs. In particular, videos are composed of content (i.e., appearance) and complex motion components that require a special mechanism to disentangle and control. To achieve this, VidStyleODE encodes the video content in a pre-trained StyleGAN W+ space and benefits from a latent ODE component to summarize the spatiotemporal dynamics of the input video. Our novel continuous video generation process then combines the two to generate high-quality and temporally consistent videos with varying frame rates. We show that our proposed method enables a variety of applications on real videos: text-guided appearance manipulation, motion manipulation, image animation, and video interpolation and extrapolation.
****
SEFD: Learning to Distill Complex Pose and Occlusion
ChangHee Yang, Kyeongbo Kong, SungJun Min, Dongyoon Wee, Ho-Deok Jang, Geonho Cha, SukJu Kang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14941-14952
This paper addresses the problem of three-dimensional (3D) human mesh estimation in complex poses and occluded situations. Although many improvements have been made in 3D human mesh estimation using the two-dimensional (2D) pose with occlusion between humans, occlusion from complex poses and other objects remains a consistent problem. Therefore, we propose the novel Skinned Multi-Person Linear (SMPL) Edge Feature Distillation (SEFD) that demonstrates robustness to complex poses and occlusions, without increasing the number of parameters compared to the baseline model. The model generates an SMPL overlapping edge similar to the ground truth that contains target person boundary and occlusion information, performing subsequent feature distillation in a simple edge map. We also perform experiments on various benchmarks and exhibit fidelity both qualitatively and quantitatively. Extensive experiments prove that our method outperforms the state-of-the-art method by 2.8% in MPJPE and 1.9% in MPVPE on a benchmark 3DPW dataset in the presence of domain gap. Also, our method is superior in 3DPW-OCC, 3DPW-PC, RH-Dataset, OCHuman, CrowdPose, and LSP dataset in which occlusion, complex pose, and domain gap exist. The code and occlusion & complex pose annotation will be available at https: //anonymous.4open.science/r/SEFD-B7F8/
****
CiT: Curation in Training for Effective Vision-Language Data
Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, Christoph Feichtenhofer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15180-15189
Large vision-language models are generally applicable to many downstream tasks, but come at an exorbitant training cost that only large institutions can afford. This paper trades generality for efficiency and presents Curation in Training (CiT), a simple and efficient vision-text learning algorithm that couples a data objective into training. CiT automatically yields quality data to speed-up contrastive image-text training and alleviates the need for an offline data filtering pipeline, allowing broad data sources (including raw image-text pairs from the web). CiT contains two loops: an outer loop curating the training data and an inner loop consuming the curated training data. The text encoder connects the two loops. Given metadata for tasks of interest, e.g., class names, and a large pool of image-text pairs, CiT alternatively selects relevant training data from the pool by measuring the similarity of their text embeddings and embeddings of the metadata. In our experiments, we observe that CiT can speed up training by over an order of magnitude, especially if the raw data size is large.
****

SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9065-9076

Neural Radiance Field (NeRF) significantly degrades when only a limited number of views are available. To complement the lack of 3D information, depth-based models, such as DSNeRF and MonoSDF, explicitly assume the availability of accurate depth maps of multiple views. They linearly scale the accurate depth maps as supervision to guide the predicted depth of few-shot NeRFs. However, accurate depth maps are difficult and expensive to capture due to wide-range depth distances in the wild.

This work presents a new Sparse-view NeRF (SparseNeRF) framework that exploits depth priors from real-world inaccurate observations. The inaccurate depth observations are either from pre-trained depth models or coarse depth maps of consumer-level depth sensors. Since coarse depth maps are not strictly scaled to the ground-truth depth maps, we propose a simple yet effective constraint, a local depth ranking method, on NeRFs such that the expected depth ranking of the NeRF is consistent with that of the coarse depth maps in local patches. To preserve the spatial continuity of the estimated depth of NeRF, we further propose a spatial continuity constraint to encourage the consistency of the expected depth continuity of NeRF with coarse depth maps. Surprisingly, with simple depth ranking constraints, SparseNeRF outperforms all state-of-the-art few-shot NeRF methods (including depth-based models) on standard LLFF and DTU datasets. Moreover, we collect a new dataset NVS-RGBD that contains real-world depth maps from Azure Kinect, ZED 2, and iPhone 13 Pro. Extensive experiments on NVS-RGBD dataset also validate the superiority and generalizability of SparseNeRF. Code and dataset are available at https://sparsenerf.github.io/.

****

Towards Models that Can See and Read

Roy Ganz, Oren Nuriel, Aviad Aberdam, Yair Kittenplon, Shai Mazor, Ron Litman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21718-21728

Visual Question Answering (VQA) and Image Captioning (CAP), which are among the most popular vision-language tasks,
have analogous scene-text versions that require reasoning from the text in the image. Despite their obvious resemblance, the two are treated independently and, as we show, yield task-specific methods that can either see or read, but not both. In this work, we conduct an in-depth analysis of this phenomenon and propose UniTNT, a Unified Text-Non-Text approach, which grants existing multimodal architectures scene-text understanding capabilities. Specifically, we treat scene-text information as an additional modality, fusing it with any pretrained encoder-decoder-based architecture via designated modules. Thorough experiments reveal that UniTNT leads to the first single model that successfully handles both task types. Moreover, we show that scene-text understanding capabilities can boost vision-language models' performance on general VQA and CAP by up to 2.69% and 0.6 CIDEr, respectively.

****

ProPainter: Improving Propagation and Transformer for Video Inpainting

Shangchen Zhou, Chongyi Li, Kelvin C.K. Chan, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10477-10486

Flow-based propagation and spatiotemporal Transformer are two mainstream mechanisms in video inpainting (VI). Despite the effectiveness of these components, they still suffer from some limitations that affect their performance. Previous propagation-based approaches are performed separately either in the image or feature domain. Global image propagation isolated from learning may cause spatial misalignment due to inaccurate optical flow. Moreover, memory or computational constraints limit the temporal range of feature propagation and video Transformer, preventing exploration of correspondence information from distant frames. To address these issues, we propose an improved framework, called ProPainter, which invo

lves enhanced ProPagation and an efficient Transformer. Specifically, we introduce dual-domain propagation that combines the advantages of image and feature warping, exploiting global correspondences reliably. We also propose a mask-guided sparse video Transformer, which achieves high efficiency by discarding unnecessary and redundant tokens. With these components, ProPainter outperforms prior arts by a large margin of 1.46 dB in PSNR while maintaining appealing efficiency.
****
Query Refinement Transformer for 3D Instance Segmentation
Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, Tianzhu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18516-18526
3D instance segmentation aims to predict a set of object instances in a scene and represent them as binary foreground masks with corresponding semantic labels. However, object instances are diverse in shape and category,and point clouds are usually sparse, unordered, and irregular, which leads to a query sampling dilemma. Besides,noise background queries interfere with proper scene perception and accurate instance segmentation. To address the above issues, we propose a Query Refinement Transformer termed QueryFormer. The key to our approach is to exploit a query initialization module to optimize the initialization process for the query distribution with a high coverage and low repetition rate. Additionally, we design an affiliated transformer decoder that suppresses the interference of noise background queries and helps the foreground queries focus on instance discriminative parts to predict final segmentation results. Extensive experiments on ScanNetV2 and S3DIS datasets show that our QueryFormer can surpass state-of-the-art 3D instance segmentation methods.
****
Root Pose Decomposition Towards Generic Non-rigid 3D Reconstruction with Monocular Videos
Yikai Wang, Yinpeng Dong, Fuchun Sun, Xiao Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13890-13900
This work focuses on the 3D reconstruction of non-rigid objects based on monocular RGB video sequences. Concretely, we aim at building high-fidelity models for generic object categories and casually captured scenes. To this end, we do not assume known root poses of objects, and do not utilize category-specific templates or dense pose priors. The key idea of our method, Root Pose Decomposition (RPD), is to maintain a per-frame root pose transformation, meanwhile building a dense field with local transformations to rectify the root pose. The optimization of local transformations is performed by point registration to the canonical space. We also adapt RPD to multi-object scenarios with object occlusions and individual differences. As a result, RPD allows non-rigid 3D reconstruction for complicated scenarios containing objects with large deformations, complex motion patterns, occlusions, and scale diversities of different individuals. Such a pipeline potentially scales to diverse sets of objects in the wild. We experimentally show that RPD surpasses state-of-the-art methods on the challenging DAVIS, OVIS, and AMA datasets. We provide video results in https://rpd-share.github.io.
****
3DHumanGAN: 3D-Aware Human Image Generation with 3D Pose Mapping
Zhuoqian Yang, Shikai Li, Wayne Wu, Bo Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23008-23019
We present 3DHumanGAN, a 3D-aware generative adversarial network that synthesizes photorealistic images of full-body humans with consistent appearances under different view-angles and body-poses. To tackle the representational and computational challenges in synthesizing the articulated structure of human bodies, we propose a novel generator architecture in which a 2D convolutional backbone is modulated by a 3D pose mapping network. The 3D pose mapping network is formulated as a renderable implicit function conditioned on a posed 3D human mesh. This design has several merits: i) it leverages the strength of 2D GANs to produce high-quality images; ii) it generates consistent images under varying view-angles and poses; iii) the model can incorporate the 3D human prior and enable pose conditioning. Project page: https://3dhumangan.github.io/.

****
LeaF: Learning Frames for 4D Point Cloud Sequence Understanding
Yunze Liu, Junyu Chen, Zekai Zhang, Jingwei Huang, Li Yi; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 604-613
We focus on learning descriptive geometry and motion features from 4D point clou
d sequences in this work. Existing works usually develop generic 4D learning too
ls without leveraging the prior that a 4D sequence comes from a single 3D scene
with local dynamics. Based on this observation, we propose to learn region-wise
coordinate frames that transform together with the underlying geometry. With suc
h frames, we can factorize geometry and motion to facilitate a feature-space geo
metric reconstruction for more effective 4D learning. To learn such region frame
s, we develop a rotation equivariant network with a frame stabilization strategy
. To leverage such frames for better spatial-temporal feature learning, we devel
op a frame-guided 4D learning scheme. Experiments show that this approach signif
icantly outperforms previous state-of-the-art methods on a wide range of 4D unde
rstanding benchmarks.
****
GLA-GCN: Global-local Adaptive Graph Convolutional Network for 3D Human Pose Est
imation from Monocular Video
Bruce X.B. Yu, Zhi Zhang, Yongxu Liu, Sheng-hua Zhong, Yan Liu, Chang Wen Chen;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2023, pp. 8818-8829
3D human pose estimation has been researched for decades with promising fruits.
3D human pose lifting is one of the promising research directions toward the tas
k where both estimated pose and ground truth pose data are used for training. Ex
isting pose lifting works mainly focus on improving the performance of estimated
 pose, but they usually underperform when testing on the ground truth pose data.
 We observe that the performance of the estimated pose can be easily improved by
 preparing good quality 2D pose, such as fine-tuning the 2D pose or using advanc
ed 2D pose detectors. As such, we concentrate on improving the 3D human pose lif
ting via ground truth data for the future improvement of more quality estimated
pose data.
 Towards this goal, a simple yet effective model called Global-local Adaptive Gr
aph Convolutional Network (GLA-GCN) is proposed in this work. Our GLA-GCN global
ly models the spatiotemporal structure via a graph representation and backtraces
 local joint features for 3D human pose estimation via individually connected la
yers.
 To validate our model design, we conduct extensive experiments on three benchma
rk datasets: Human3.6M, HumanEva-I, and MPI-INF-3DHP. Experimental results show
that our GLA-GCN implemented with ground truth 2D poses significantly outperform
s state-of-the-art methods (e.g., up to 3%, 17%, and 14% error reductions on Hum
an3.6M, HumanEva-I, and MPI-INF-3DHP, respectively).
****
Snow Removal in Video: A New Dataset and A Novel Method
Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, Lei Zh
u; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 13211-13222
Snowfall is a common weather phenomenon that can severely affect computer vision
 tasks by obscuring objects and scenes. However, existing deep learning-based sn
ow removal methods are designed for single images only. In this paper, we target
 a more complex task -- video snow removal, which aims to restore the clear vide
o from the snowy video. To facilitate this task, we propose the first high-quali
ty video dataset, which simulates realistic physical characteristics of snow and
 haze using a rendering engine and augmentation techniques. We also develop a de
ep learning framework for video snow removal. It involves Specifically, we propo
se a snow-query temporal aggregation module and a snow-aware contrastive learnin
g loss function. The module aggregates features between video frames and removes
 snow effectively, while the loss function helps identify and eliminate snow fea
tures. We conduct extensive experiments and demonstrate that our proposed datase
t is more realistic than previous datasets, and the models trained on it achieve

better performance in real-world snowing images. Our proposed method outperforms state-of-the-art video and image-based methods on both synthetic and real snowy videos.
****
Degradation-Resistant Unfolding Network for Heterogeneous Image Fusion
Chunming He, Kai Li, Guoxia Xu, Yulun Zhang, Runze Hu, Zhenhua Guo, Xiu Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12611-12621
Heterogeneous image fusion (HIF) techniques aim to enhance image quality by merging complementary information from images captured by different sensors. Among these algorithms, deep unfolding network (DUN)-based methods achieve promising performance but still suffer from two issues: they lack a degradation-resistant-oriented fusion model and struggle to adequately consider the structural properties of DUNs, making them vulnerable to degradation scenarios. In this paper, we propose a Degradation-Resistant Unfolding Network (DeRUN) for the HIF task to generate high-quality fused images even in degradation scenarios. Specifically, we introduce a novel HIF model for degradation resistance and derive its optimization procedures. Then, we incorporate the optimization unfolding process into the proposed DeRUN for end-to-end training. To ensure the robustness and efficiency of DeRUN, we employ a joint constraint strategy and a lightweight partial weight sharing module. To train DeRUN, we further propose a gradient direction-based entropy loss with powerful texture representation capacity. Extensive experiments show that DeRUN significantly outperforms existing methods on four HIF tasks, as well as downstream applications, with cheaper computational and memory costs.
****
Priority-Centric Human Motion Generation in Discrete Latent Space
Hanyang Kong, Kehong Gong, Dongze Lian, Michael Bi Mi, Xinchao Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14806-14816
Text-to-motion generation is a formidable task, aiming to produce human motions that align with the input text while also adhering to human capabilities and physical laws. While there have been advancements in diffusion models, their application in discrete spaces remains underexplored. Current methods often overlook the varying significance of different motions, treating them uniformly. It is essential to recognize that not all motions hold the same relevance to a particular textual description. Some motions, being more salient and informative, should be given precedence during generation. In response, we introduce a Priority-Centric Motion Discrete Diffusion Model (M2DM), which utilizes a Transformer-based VQ-VAE to derive a concise, discrete motion representation, incorporating a global self-attention mechanism and a regularization term to counteract code collapse. We also present a motion discrete diffusion model that employs an innovative noise schedule, determined by the significance of each motion token within the entire motion sequence. This approach retains the most salient motions during the reverse diffusion process, leading to more semantically rich and varied motions. Additionally, we formulate two strategies to gauge the importance of motion tokens, drawing from both textual and visual indicators. Comprehensive experiments on the HumanML3D and KIT-ML datasets confirm that our model surpasses existing techniques in fidelity and diversity, particularly for intricate textual descriptions.
****
Domain-Specificity Inducing Transformers for Source-Free Domain Adaptation
Sunandini Sanyal, Ashish Ramayee Asokan, Suvaansh Bhambri, Akshay Kulkarni, Jogendra Nath Kundu, R Venkatesh Babu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18928-18937
Conventional Domain Adaptation (DA) methods aim to learn domain-invariant feature representations to improve the target adaptation performance. However, we motivate that domain-specificity is equally important since in-domain trained models hold crucial domain-specific properties that are beneficial for adaptation. Hence, we propose to build a framework that supports disentanglement and learning of domain-specific factors and task-specific factors in a unified model. Motivate

d by the success of vision transformers in several multi-modal vision problems, we find that queries could be leveraged to extract the domain-specific factors. Hence, we propose a novel Domain-Specificity inducing Transformer (DSiT) framework for disentangling and learning both domain-specific and task-specific factors. To achieve disentanglement, we propose to construct novel Domain-Representative Inputs (DRI) with domain-specific information to train a domain classifier with a novel domain token. We are the first to utilize vision transformers for domain adaptation in a privacy-oriented source-free setting, and our approach achieves state-of-the-art performance on single-source, multi-source, and multi-target benchmarks.
****
Towards Improved Input Masking for Convolutional Neural Networks
Sriram Balasubramanian, Soheil Feizi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1855-1865
The ability to remove features from the input of machine learning models is very important to understand and interpret model predictions. However, this is non-trivial for vision models since masking out parts of the input image typically causes large distribution shifts. This is because the baseline color used for masking (typically grey or black) is out of distribution. Furthermore, the shape of the mask itself can contain unwanted signals which can be used by the model for its predictions. Recently, there has been some progress in mitigating this issue (called missingness bias) in image masking for vision transformers. In this work, we propose a new masking method for CNNs we call layer masking in which the missingness bias caused by masking is reduced to a large extent. Intuitively, layer masking applies a mask to intermediate activation maps so that the model only processes the unmasked input. We show that our method (i) is able to eliminate or minimize the influence of the mask shape or color on the output of the model, and (ii) is much better than replacing the masked region by black or grey for input perturbation based interpretability techniques like LIME. Thus, layer masking is much less affected by missingness bias than other masking strategies. We also demonstrate how the shape of the mask may leak information about the class, thus affecting estimates of model reliance on class-relevant features derived from input masking. Furthermore, we discuss the role of data augmentation techniques for tackling this problem, and argue that they are not sufficient for preventing model reliance on mask shape.
****
3DHacker: Spectrum-based Decision Boundary Generation for Hard-label 3D Point Cloud Attack
Yunbo Tao, Daizong Liu, Pan Zhou, Yulai Xie, Wei Du, Wei Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14340-14350
With the maturity of depth sensors, the vulnerability of 3D point cloud models has received increasing attention in various applications such as autonomous driving and robot navigation. Previous 3D adversarial attackers either follow the white-box setting to iteratively update the coordinate perturbations based on gradients, or utilize the output model logits to estimate noisy gradients in the black-box setting. However, these attack methods are hard to be deployed in real-world scenarios since realistic 3D applications will not share any model details to users. Therefore, we explore a more challenging yet practical 3D attack setting, i.e., attacking point clouds with black-box hard labels, in which the attacker can only have access to the prediction label of the input. To tackle this setting, we propose a novel 3D attack method, termed 3D Hard-label attacker (3DHacker), based on the developed decision boundary algorithm to generate adversarial samples solely with the knowledge of class labels. Specifically, to construct the class-aware model decision boundary, 3DHacker first randomly fuses two point clouds of different classes in the spectral domain to craft their intermediate sample with high imperceptibility, then projects it onto the decision boundary via binary search. To restrict the final perturbation size, 3DHacker further introduces an iterative optimization strategy to move the intermediate sample along the decision boundary for generating adversarial point clouds with smallest trivial

perturbations. Extensive evaluations show that, even in the challenging hard-label setting, 3DHacker still competitively outperforms existing 3D attacks regarding the attack performance as well as adversary quality.
****
Exploring Lightweight Hierarchical Vision Transformers for Efficient Visual Tracking
Ben Kang, Xin Chen, Dong Wang, Houwen Peng, Huchuan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9612-9621
Transformer-based visual trackers have demonstrated significant progress owing to their superior modeling capabilities. However, existing trackers are hampered by low speed, limiting their applicability on devices with limited computational power. To alleviate this problem, we propose HiT, a new family of efficient tracking models that can run at high speed on different devices while retaining high performance. The central idea of HiT is the Bridge Module, which bridges the gap between modern lightweight transformers and the tracking framework. The Bridge Module incorporates the high-level information of deep features into the shallow large-resolution features. In this way, it produces better features for the tracking head. We also propose a novel dual-image position encoding technique that simultaneously encodes the position information of both the search region and template images. The HiT model achieves promising speed with competitive performance. For instance, it runs at 61 frames per second (fps) on the Nvidia Jetson AGX edge device. Furthermore, HiT attains 64.6% AUC on the LaSOT benchmark, surpassing all previous efficient trackers.
****
Improving Zero-Shot Generalization for CLIP with Synthesized Prompts
Zhengbo Wang, Jian Liang, Ran He, Nan Xu, Zilei Wang, Tieniu Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3032-3042
With the growing interest in pretrained vision-language models like CLIP, recent research has focused on adapting these models to downstream tasks. Despite achieving promising results, most existing methods require labeled data for all classes, which may not hold in real-world applications due to the long tail and Zipf's law. For example, some classes may lack labeled data entirely, such as emerging concepts. To address this problem, we propose a plug-and-play generative approach called Synt\HesIzed Prompts (SHIP) to improve existing fine-tuning methods. Specifically, we follow variational autoencoders to introduce a generator that reconstructs the visual features by inputting the synthesized prompts and the corresponding class names to the textual encoder of CLIP. In this manner, we easily obtain the synthesized features for the remaining label-only classes. Thereafter, we fine-tune CLIP with off-the-shelf methods by combining labeled and synthesized features. Extensive experiments on base-to-new generalization, cross-dataset transfer learning, and generalized zero-shot learning demonstrate the superiority of our approach.
****
MiniROAD: Minimal RNN Framework for Online Action Detection
Joungbin An, Hyolim Kang, Su Ho Han, Ming-Hsuan Yang, Seon Joo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10341-10350
Online Action Detection (OAD) is the task of identifying actions in streaming videos without access to future frames. Much effort has been devoted to effectively capturing long-range dependencies, with transformers receiving the spotlight for their ability to capture long-range temporal structures. In contrast, RNNs have received less attention lately, due to their lower performance compared to recent methods that utilize transformers. In this paper, we investigate the underlying reasons for the inferior performance of RNNs compared to transformer-based algorithms. Our findings indicate that the discrepancy between training and inference is the primary hindrance to the effective training of RNNs. To address this, we propose applying non-uniform weights to the loss computed at each time step, which allows the RNN model to learn from the predictions made in an environment that better resembles the inference stage. Extensive experiments on three ben

chmark datasets, THUMOS, TVSeries, and FineAction demonstrate that a minimal RNN-based model trained with the proposed methodology performs equally or better than the existing best methods with a significant increase in efficiency. The code is available at https://github.com/jbistanbul/MiniROAD.
****

Efficient Emotional Adaptation for Audio-Driven Talking-Head Generation

Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22634-22645

Audio-driven talking-head synthesis is a popular research topic for virtual human-related applications. However, the inflexibility and inefficiency of existing methods, which necessitate expensive end-to-end training to transfer emotions from guidance videos to talking-head predictions, are significant limitations. In this work, we propose the Emotional Adaptation for Audio-driven Talking-head (EAT) method, which transforms emotion-agnostic talking-head models into emotion-controllable ones in a cost-effective and efficient manner through parameter-efficient adaptations. Our approach utilizes a pretrained emotion-agnostic talking-head transformer and introduces three lightweight adaptations (the Deep Emotional Prompts, Emotional Deformation Network, and Emotional Adaptation Module) from different perspectives to enable precise and realistic emotion controls. Our experiments demonstrate that our approach achieves state-of-the-art performance on widely-used benchmarks, including LRW and MEAD. Additionally, our parameter-efficient adaptations exhibit remarkable generalization ability, even in scenarios where emotional training videos are scarce or nonexistent. Project website: https://yuangan.github.io/eat/
****

Object-aware Gaze Target Detection

Francesco Tonini, Nicola Dall'Asen, Cigdem Beyan, Elisa Ricci; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21860-21869

Gaze target detection aims to predict the image location where the person is looking and the probability that a gaze is out of the scene. Several works have tackled this task by regressing a gaze heatmap centered on the gaze location, however, they overlooked decoding the relationship between the people and the gazed objects. This paper proposes a Transformer-based architecture that automatically detects objects (including heads) in the scene to build associations between every head and the gazed-head/object, resulting in a comprehensive, explainable gaze analysis composed of: gaze target area, gaze pixel point, the class and the image location of the gazed-object. Upon evaluation of the in-the-wild benchmarks, our method achieves state-of-the-art results on all metrics (up to 2.91% gain in AUC, 50% reduction in gaze distance, and 9% gain in out-of-frame average precision) for gaze target detection and 11-13% improvement in average precision for the classification and the localization of the gazed-objects. The code of the proposed method is publicly available.
****

Gramian Attention Heads are Strong yet Efficient Vision Learners

Jongbin Ryu, Dongyoon Han, Jongwoo Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5841-5851

We introduce a novel architecture design that enhances expressiveness by incorporating multiple head classifiers (i.e., classification heads) instead of relying on channel expansion or additional building blocks. Our approach employs attention-based aggregation, utilizing pairwise feature similarity to enhance multiple lightweight heads with minimal resource overhead. We compute the Gramian matrices to reinforce class tokens in an attention layer for each head. This enables the heads to learn more discriminative representations, enhancing their aggregation capabilities. Furthermore, we propose a learning algorithm that encourages heads to complement each other by reducing correlation for aggregation. Our models eventually surpass state-of-the-art CNNs and ViTs regarding the accuracy-throughput trade-off on ImageNet-1K and deliver remarkable performance across various downstream tasks, such as COCO object instance segmentation, ADE20k semantic segmentation, and fine-grained visual classification datasets. The effectiveness of

our framework is substantiated by practical experimental results and further underpinned by generalization error bound. We release the code publicly at: https://github.com/Lab-LVM/imagenet-models.
****

VADER: Video Alignment Differencing and Retrieval
Alexander Black, Simon Jenni, Tu Bui, Md. Mehrab Tanjim, Stefano Petrangeli, Ritwik Sinha, Viswanathan Swaminathan, John Collomosse; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22357-22367
We propose VADER, a spatio-temporal matching, alignment, and change summarization method to help fight misinformation spread via manipulated videos. VADER matches and coarsely aligns partial video fragments to candidate videos using a robust visual descriptor and scalable search over adaptively chunked video content. A transformer-based alignment module then refines the temporal localization of the query fragment within the matched video. A space-time comparator module identifies regions of manipulation between aligned content, invariant to any changes due to any residual temporal misalignments or artifacts arising from non-editorial changes of the content. Robustly matching video to a trusted source enables conclusions to be drawn on video provenance, enabling informed trust decisions on content encountered. Code and data are available at https://github.com/AlexBlck/vader
****

MI-GAN: A Simple Baseline for Image Inpainting on Mobile Devices
Andranik Sargsyan, Shant Navasardyan, Xingqian Xu, Humphrey Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7335-7345
In recent years, many deep learning based image inpainting methods have been developed by the research community. Some of those methods have shown impressive image completion abilities. Yet, to the best of our knowledge, there is no image inpainting model designed to run on mobile devices. In this paper we present a simple image inpainting baseline, Mobile Inpainting GAN (MI-GAN), which is approximately one order of magnitude computationally cheaper and smaller than existing state-of-the-art inpainting models, and can be efficiently deployed on mobile devices. Excessive quantitative and qualitative evaluations show that MI-GAN performs comparable or, in some cases, better than recent state-of-the-art approaches. Moreover, we perform a user study comparing MI-GAN results with results from several commercial mobile inpainting applications, which clearly shows the advantage of MI-GAN in comparison to existing apps. With the purpose of high quality and efficient inpainting, we utilize an effective combination of adversarial training, model re-parametrization, and knowledge distillation. Our models and code are publicly available at https://github.com/Picsart-AI-Research/MI-GAN.
****

HiLo: Exploiting High Low Frequency Relations for Unbiased Panoptic Scene Graph Generation
Zijian Zhou, Miaojing Shi, Holger Caesar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21637-21648
Panoptic Scene Graph generation (PSG) is a recently proposed task in image scene understanding that aims to segment the image and extract triplets of subjects, objects and their relations to build a scene graph. This task is particularly challenging for two reasons. First, it suffers from a long-tail problem in its relation categories, making naive biased methods more inclined to high-frequency relations. Existing unbiased methods tackle the long-tail problem by data/loss rebalancing to favor low-frequency relations. Second, a subject-object pair can have two or more semantically overlapping relations. While existing methods favor one over the other, our proposed HiLo framework lets different network branches specialize on low and high frequency relations, enforce their consistency and fuse the results. To the best of our knowledge we are the first to propose an explicitly unbiased PSG method. In extensive experiments we show that our HiLo framework achieves state-of-the-art results on the PSG task. We also apply our method to the Scene Graph Generation task that predicts boxes instead of masks and see improvements over all baseline methods. Code is available at https://github.com/

franciszzj/HiLo.
****

Chop & Learn: Recognizing and Generating Object-State Compositions
Nirat Saini, Hanyu Wang, Archana Swaminathan, Vinoj Jayasundara, Bo He, Kamal Gupta, Abhinav Shrivastava; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20247-20258
Recognizing and generating object-state compositions has been a challenging task, especially when generalizing to unseen compositions. In this paper, we study the task of cutting objects in different styles and the resulting object state changes. We propose a new benchmark suite Chop & Learn, to accommodate the needs of learning objects and different cut styles using multiple viewpoints. We also propose a new task of Compositional Image Generation, which can transfer learned cut styles to different objects, by generating novel object-state images. Moreover, we also use the videos for Compositional Action Recognition, and show valuable uses of this dataset for multiple video tasks. Project website: https://chopnlearn.github.io.
****

Automatic Animation of Hair Blowing in Still Portrait Photos
Wenpeng Xiao, Wentao Liu, Yitong Wang, Bernard Ghanem, Bing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22963-22975
We propose a novel approach to animate human hair in a still portrait photo. Existing work has largely studied the animation of fluid elements such as water and fire. However, hair animation for a real image remains underexplored, which is a challenging problem, due to the high complexity of hair structure and dynamics. Considering the complexity of hair structure, we innovatively treat hair wisp extraction as an instance segmentation problem, where a hair wisp is referred to as an instance. With advanced instance segmentation networks, our method extracts meaningful and natural hair wisps. Furthermore, we propose a wisp-aware animation module that animates hair wisps with pleasing motions without noticeable artifacts. The extensive experiments show the superiority of our method. Our method provides the most pleasing and compelling viewing experience in the qualitative experiments, and outperforms state-of-the-art still-image animation methods by a large margin in the quantitative evaluation. Project url: https://nevergiveu.github.io/AutomaticHairBlowing/
****

A Large-Scale Outdoor Multi-Modal Dataset and Benchmark for Novel View Synthesis and Implicit Scene Reconstruction
Chongshan Lu, Fukun Yin, Xin Chen, Wen Liu, Tao Chen, Gang Yu, Jiayuan Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7557-7567
Neural Radiance Fields (NeRF) has achieved impressive results in single object scene reconstruction and novel view synthesis, as demonstrated on many single modality and single object focused indoor scene datasets like DTU, BMVS, and NeRF Synthetic. However, the study of NeRF on large-scale outdoor scene reconstruction is still limited, as there is no unified outdoor scene dataset for large-scale NeRF evaluation due to expensive data acquisition and calibration costs.
  In this work, we propose a large-scale outdoor multi-modal dataset, OMMO dataset, containing complex objects and scenes with calibrated images, point clouds and prompt annotations. A new benchmark for several outdoor NeRF-based tasks is established, such as novel view synthesis,diverse 3D representation, and multi-modal NeRF. To create the dataset, we capture and collect a large number of real fly-view videos and select high-quality and high-resolution clips from them. Then we design a quality review module to refine images, remove low-quality frames and fail-to-calibrate scenes through a learning-based automatic evaluation plus manual review. Finally, volunteers are employed to label and review the prompt annotation for each scene and keyframe.Compared with existing NeRF datasets, our dataset contains abundant real-world urban and natural scenes with various scales, camera trajectories, and lighting conditions. Experiments show that our dataset can benchmark most state-of-the-art NeRF methods on different tasks.The datase

t can be found at the following link: https://ommo.luchongshan.com/ .
****
4D Panoptic Segmentation as Invariant and Equivariant Field Prediction
Minghan Zhu, Shizhong Han, Hong Cai, Shubhankar Borse, Maani Ghaffari, Fatih Porikli; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22488-22498
In this paper, we develop rotation-equivariant neural networks for 4D panoptic segmentation. 4D panoptic segmentation is a benchmark task for autonomous driving that requires recognizing semantic classes and object instances on the road based on LiDAR scans, as well as assigning temporally consistent IDs to instances across time. We observe that the driving scenario is symmetric to rotations on the ground plane. Therefore, rotation-equivariance could provide better generalization and more robust feature learning. Specifically, we review the object instance clustering strategies and restate the centerness-based approach and the offset-based approach as the prediction of invariant scalar fields and equivariant vector fields. Other sub-tasks are also unified from this perspective, and different invariant and equivariant layers are designed to facilitate their predictions. Through evaluation on the standard 4D panoptic segmentation benchmark of SemanticKITTI, we show that our equivariant models achieve higher accuracy with lower computational costs compared to their non-equivariant counterparts. Moreover, our method sets the new state-of-the-art performance and achieves 1st place on the SemanticKITTI 4D Panoptic Segmentation leaderboard.
****
Unleashing Vanilla Vision Transformer with Masked Image Modeling for Object Detection
Yuxin Fang, Shusheng Yang, Shijie Wang, Yixiao Ge, Ying Shan, Xinggang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6244-6253
We present an approach to efficiently and effectively adapt a masked image modeling (MIM) pre-trained vanilla Vision Transformer (ViT) for object detection, which is based on our two novel observations: (i) A MIM pre-trained vanilla ViT encoder can work surprisingly well in the challenging object-level recognition scenario even with randomly sampled partial observations, e.g., only 25%   50% of the input embeddings. (ii) In order to construct multi-scale representations for object detection from single-scale ViT, a randomly initialized compact convolutional stem supplants the pre-trained patchify stem, and its intermediate features can naturally serve as the higher resolution inputs of a feature pyramid network without further upsampling or other manipulations. While the pre-trained ViT is only regarded as the third-stage of our detector's backbone instead of the whole feature extractor. This naturally results in a ConvNet-ViT hybrid architecture. The proposed detector, named MIMDet, enables a MIM pre-trained vanilla ViT to outperform leading hierarchical architectures such as Swin Transformer, MViTv2 and ConvNeXt on COCO object detection & instance segmentation, and achieves better results compared with the previous best adapted vanilla ViT detector using a more modest fine-tuning recipe while converging 2.8x faster. Code and pre-trained models are available at https://github.com/hustvl/MIMDet.
****
NDC-Scene: Boost Monocular 3D Semantic Scene Completion in Normalized Device Coordinates Space
Jiawei Yao, Chuming Li, Keqiang Sun, Yingjie Cai, Hao Li, Wanli Ouyang, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9455-9465
Monocular 3D Semantic Scene Completion (SSC) has garnered significant attention in recent years due to its potential to predict complex semantics and geometry shapes from a single image, requiring no 3D inputs. In this paper, we identify several critical issues in current state-of-the-art methods, including the Feature Ambiguity of projected 2D features in the ray to the 3D space, the Pose Ambiguity of the 3D convolution, and the Computation Imbalance in the 3D convolution across different depth levels. To address these problems, we devise a novel Normalized Device Coordinates scene completion network (NDC-Scene) that directly exten

ds the 2D feature map to a Normalized Device Coordinates (NDC) space, rather than to the world space directly, through progressive restoration of the dimension of depth with deconvolution operations. Experiment results demonstrate that transferring the majority of computation from the target 3D space to the proposed normalized device coordinates space benefits monocular SSC tasks. Additionally, we design a Depth-Adaptive Dual Decoder to simultaneously upsample and fuse the 2D and 3D feature maps, further improving overall performance. Our extensive experiments confirm that the proposed method consistently outperforms state-of-the-art methods on both outdoor SemanticKITTI and indoor NYUv2 datasets. Our code are available at https://github.com/Jiawei-Yao0812/NDCScene.
****
Spatio-Temporal Crop Aggregation for Video Representation Learning
Sepehr Sameni, Simon Jenni, Paolo Favaro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5664-5674
We propose Spatio-temporal Crop Aggregation for video representation LEarning (SCALE), a novel method that enjoys high scalability at both training and inference time. Our model builds long-range video features by learning from sets of video clip-level features extracted with a pre-trained backbone. To train the model, we propose a self-supervised objective consisting of masked clip feature predictions. We apply sparsity to both the input, by extracting a random set of video clips, and to the loss function, by only reconstructing the sparse inputs. Moreover, we use dimensionality reduction by working in the latent space of a pre-trained backbone applied to single video clips. These techniques make our method not only extremely efficient to train but also highly effective in transfer learning. We demonstrate that our video representation yields state-of-the-art performance with linear, nonlinear, and k-NN probing on common action classification and video understanding datasets.
****
Zip-NeRF: Anti-Aliased Grid-Based Neural Radiance Fields
Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, Peter Hedman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19697-19705
Neural Radiance Field training can be accelerated through the use of grid-based representations in NeRF's learned mapping from spatial coordinates to colors and volumetric density. However, these grid-based approaches lack an explicit understanding of scale and therefore often introduce aliasing, usually in the form of jaggies or missing scene content. Anti-aliasing has previously been addressed by mip-NeRF 360, which reasons about sub-volumes along a cone rather than points along a ray, but this approach is not natively compatible with current grid-based techniques. We show how ideas from rendering and signal processing can be used to construct a technique that combines mip-NeRF 360 and grid-based models such as Instant NGP to yield error rates that are 8%-77% lower than either prior technique, and that trains 24x faster than mip-NeRF 360.
****
Neural-PBIR Reconstruction of Shape, Material, and Illumination
Cheng Sun, Guangyan Cai, Zhengqin Li, Kai Yan, Cheng Zhang, Carl Marshall, Jia-Bin Huang, Shuang Zhao, Zhao Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18046-18056
Reconstructing the shape and spatially varying surface appearances of a physical-world object as well as its surrounding illumination based on 2D images (e.g., photographs) of the object has been a long-standing problem in computer vision and graphics. In this paper, we introduce an accurate and highly efficient object reconstruction pipeline combining neural based object reconstruction and physics-based inverse rendering (PBIR). Our pipeline firstly leverages a neural SDF based shape reconstruction to produce high-quality but potentially imperfect object shape. Then, we introduce a neural material and lighting distillation stage to achieve high-quality predictions for material and illumination. In the last stage, initialized by the neural predictions, we perform PBIR to refine the initial results and obtain the final high-quality reconstruction of object shape, material, and illumination. Experimental results demonstrate our pipeline significant

ly outperforms existing methods quality-wise and performance-wise. Code: https://neural-pbir.github.io/
****
Fg-T2M: Fine-Grained Text-Driven Human Motion Generation via Diffusion Model
Yin Wang, Zhiying Leng, Frederick W. B. Li, Shun-Cheng Wu, Xiaohui Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22035-22044
Text-driven human motion generation in computer vision is both significant and challenging. However, current methods are limited to producing either deterministic or imprecise motion sequences, failing to effectively control the temporal and spatial relationships required to conform to a given text description. In this work, we propose a fine-grained method for generating high-quality, conditional human motion sequences supporting precise text description. Our approach consists of two key components: 1) a linguistics-structure assisted module that constructs accurate and complete language feature to fully utilize text information; and 2) a context-aware progressive reasoning module that learns neighborhood and overall semantic linguistics features from shallow and deep graph neural networks to achieve a multi-step inference. Experiments show that our approach outperforms text-driven motion generation methods on HumanML3D and KIT test sets and generates better visually confirmed motion to the text conditions.
****
BlindHarmony: "Blind" Harmonization for MR Images via Flow Model
Hwihun Jeong, Heejoon Byun, Dong Un Kang, Jongho Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21129-21139
In MRI, images of the same contrast (e.g., T1) from the same subject can exhibit noticeable differences when acquired using different hardware, sequences, or scan parameters. These differences in images create a domain gap that needs to be bridged by a step called image harmonization, to process the images successfully using conventional or deep learning-based image analysis (e.g., segmentation). Several methods, including deep learning-based approaches, have been proposed to achieve image harmonization. However, they often require datasets from multiple domains for deep learning training and may still be unsuccessful when applied to images from unseen domains. To address this limitation, we propose a novel concept called 'Blind Harmonization', which utilizes only target domain data for training but still has the capability to harmonize images from unseen domains. For the implementation of blind harmonization, we developed BlindHarmony using an unconditional flow model trained on target domain data. The harmonized image is optimized to have a correlation with the input source domain image while ensuring that the latent vector of the flow model is close to the center of the Gaussian distribution. BlindHarmony was evaluated on both simulated and real datasets and compared to conventional methods. BlindHarmony demonstrated noticeable performance on both datasets, highlighting its potential for future use in clinical settings. The source code is available at: https://github.com/SNU-LIST/BlindHarmony
****
Zero-guidance Segmentation Using Zero Segment Labels
Pitchaporn Rewatbowornwong, Nattanat Chatthee, Ekapol Chuangsuwanich, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1162-1172
The joint visual-language model CLIP has enabled new and exciting applications, such as open-vocabulary segmentation, which can locate any segment given an arbitrary text query. In our research, we ask whether it is possible to discover semantic segments without any user guidance in the form of text queries or predefined classes, and label them using natural language automatically? We propose a novel problem zero-guidance segmentation and the first baseline that leverages two pre-trained generalist models, DINO and CLIP, to solve this problem without any fine-tuning or segmentation dataset. The general idea is to first segment an image into small over-segments, encode them into CLIP's visual-language space, translate them into text labels, and merge semantically similar segments together. The key challenge, however, is how to encode a visual segment into a segment-specific embedding that balances global and local context information, both useful

for recognition. Our main contribution is a novel attention-masking technique th at balances the two contexts by analyzing the attention layers inside CLIP. We a lso introduce several metrics for the evaluation of this new task. With CLIP's i nnate knowledge, our method can precisely locate the Mona Lisa painting among a museum crowd.
****
Efficient LiDAR Point Cloud Oversegmentation Network
Le Hui, Linghua Tang, Yuchao Dai, Jin Xie, Jian Yang; Proceedings of the IEEE/CV F International Conference on Computer Vision (ICCV), 2023, pp. 18003-18012
Point cloud oversegmentation is a challenging task since it needs to produce per ceptually meaningful partitions (i.e., superpoints) of a point cloud. Most exist ing oversegmentation methods cannot efficiently generate superpoints from large-scale LiDAR point clouds due to complex and inefficient procedures. In this pape r, we propose a simple yet efficient end-to-end LiDAR oversegmentation network, which segments superpoints from the LiDAR point cloud by grouping points based o n low-level point embeddings. Specifically, we first learn the similarity of poi nts from the constructed local neighborhoods to obtain low-level point embedding s through the local discriminative loss. Then, to generate homogeneous superpoin ts from the sparse LiDAR point cloud, we propose a LiDAR point grouping algorith m that simultaneously considers the similarity of point embeddings and the Eucli dean distance of points in 3D space. Finally, we design a superpoint refinement module for accurately assigning the hard boundary points to the corresponding su perpoints. Extensive results on two large-scale outdoor datasets, SemanticKITTI and nuScenes, show that our method achieves a new state-of-the-art in LiDAR over segmentation. Notably, the inference time of our method is 100x faster than that of other methods. Furthermore, we apply the learned superpoints to the LiDAR se mantic segmentation task and the results show that using superpoints can signifi cantly improve the LiDAR semantic segmentation of the baseline network. Code is available at https://github.com/fpthink/SuperLiDAR.
****
Communication-efficient Federated Learning with Single-Step Synthetic Features C ompressor for Faster Convergence
Yuhao Zhou, Mingjia Shi, Yuanxi Li, Yanan Sun, Qing Ye, Jiancheng Lv; Proceeding s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5031-5040
Reducing communication overhead in federated learning (FL) is challenging but cr ucial for large-scale distributed privacy-preserving machine learning. While met hods utilizing sparsification or other techniques can largely reduce the communi cation overhead, the convergence rate is also greatly compromised. In this paper , we propose a novel method named Single-Step Synthetic Features Compressor (3SF C) to achieve communication-efficient FL by directly constructing a tiny synthet ic dataset containing synthetic features based on raw gradients. Therefore, 3SFC can achieve an extremely low compression rate when the constructed synthetic da taset contains only one data sample. Additionally, the compressing phase of 3SFC utilizes a similarity-based objective function so that it can be optimized with just one step, considerably improving its performance and robustness. To minimi ze the compressing error, error feedback (EF) is also incorporated into 3SFC. Ex periments on multiple datasets and models suggest that 3SFC has significantly be tter convergence rates compared to competing methods with lower compression rate s (i.e., up to 0.02%). Furthermore, ablation studies and visualizations show tha t 3SFC can carry more information than competing methods for every communication round, further validating its effectiveness.
****
SVDFormer: Complementing Point Cloud via Self-view Augmentation and Self-structu re Dual-generator
Zhe Zhu, Honghua Chen, Xing He, Weiming Wang, Jing Qin, Mingqiang Wei; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14508-14518
In this paper, we propose a novel network, SVDFormer, to tackle two specific cha llenges in point cloud completion: understanding faithful global shapes from inc

omplete point clouds and generating high-accuracy local structures. Current methods either perceive shape patterns using only 3D coordinates or import extra images with well-calibrated intrinsic parameters to guide the geometry estimation of the missing parts. However, these approaches do not always fully leverage the cross-modal self-structures available for accurate and high-quality point cloud completion. To this end, we first design a Self-view Fusion Network that leverages multiple-view depth image information to observe incomplete self-shape and generate a compact global shape. To reveal highly detailed structures, we then introduce a refinement module, called Self-structure Dual-generator, in which we incorporate learned shape priors and geometric self-similarities for producing new points. By perceiving the incompleteness of each point, the dual-path design disentangles refinement strategies conditioned on the structural type of each point. SVDFormer absorbs the wisdom of self-structures, avoiding any additional paired information such as color images with precisely calibrated camera intrinsic parameters. Comprehensive experiments indicate that our method achieves state-of-the-art performance on widely-used benchmarks. Code is available at https://github.com/czvvd/SVDFormer.
****
Few-Shot Video Classification via Representation Fusion and Promotion Learning
Haifeng Xia, Kai Li, Martin Renqiang Min, Zhengming Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19311-19320
Recent few-shot video classification (FSVC) works achieve promising performance by capturing similarity across support and query samples with different temporal alignment strategies or learning discriminative features via Transformer block within each episode. However, they ignore two important issues: a) It is difficult to capture rich intrinsic action semantics from a limited number of support instances within each task. b) Redundant or irrelevant frames in videos easily weaken the positive influence of discriminative frames. To address these two issues, this paper proposes a novel Representation Fusion and Promotion Learning (RFPL) mechanism with two sub-modules: meta-action learning (MAL) and reinforced image representation (RIR). Concretely, during training stage, we perform online learning for seeking a task-shared meta-action bank to enrich task-specific action representation by injecting global knowledge. Besides, we exploit reinforcement learning to obtain the importance of each frame and refine the representation. This operation maximizes the contribution of discriminative frames to further capture the similarity of support and query samples from the same category. Our RFPL framework is highly flexible that it can be integrated with many existing FSVC methods. Extensive experiments show that RFPL significantly enhances the performance of existing FSVC models when integrated with them.
****
E3Sym: Leveraging E(3) Invariance for Unsupervised 3D Planar Reflective Symmetry Detection
Ren-Wu Li, Ling-Xiao Zhang, Chunpeng Li, Yu-Kun Lai, Lin Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14543-14553
Detecting symmetrical properties is a fundamental task in 3D shape analysis. In the case of a 3D model with planar symmetries, each point has a corresponding mirror point w.r.t. a symmetry plane, and the correspondences remain invariant under any arbitrary Euclidean transformation. Our proposed method, E3Sym, aims to detect planar reflective symmetry in an unsupervised and end-to-end manner by leveraging E(3) invariance. E3Sym establishes robust point correspondences through the use of E(3) invariant features extracted from a lightweight neural network, from which the dense symmetry prediction is produced. We also introduce a novel and efficient clustering algorithm to aggregate the dense prediction and produce a detected symmetry set, allowing for the detection of an arbitrary number of planar symmetries while ensuring the method remains differentiable for end-to-end training. Our method also possesses the ability to infer reasonable planar symmetries from incomplete shapes, which remains challenging for existing methods. Extensive experiments demonstrate that E3Sym is both effective and robust, outperforming state-of-the-art methods.

****
CTVIS: Consistent Training for Online Video Instance Segmentation
Kaining Ying, Qing Zhong, Weian Mao, Zhenhua Wang, Hao Chen, Lin Yuanbo Wu, Yifan Liu, Chengxiang Fan, Yunzhi Zhuge, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 899-908
The discrimination of instance embeddings plays a vital role in associating instances across time for online video instance segmentation (VIS). Instance embedding learning is directly supervised by the contrastive loss computed upon the contrastive items (CIs), which are sets of anchor/positive/negative embeddings. Recent online VIS methods leverage CIs sourced from one reference frame only, which we argue is insufficient for learning highly discriminative embeddings. Intuitively, a possible strategy to enhance CIs is replicating the inference phase during training. To this end, we propose a simple yet effective training strategy, called Consistent Training for Online VIS(CTVIS), which devotes to aligning the training and inference pipelines in terms of building CIs. Specifically, CTVIS constructs CIs by referring inference the momentum-averaged embedding and the memory bank storage mechanisms, and adding noise to the relevant embeddings. Such an extension allows a reliable comparison between embeddings of current instances and the stable representations of historical instances, thereby conferring an advantage in modeling VIS challenges such as occlusion, re-identification, and deformation. Empirically, CTVIS outstrips the SOTA VIS models by up to +5.0 points on three VIS benchmarks, including YTVIS19 (55.1% AP), YTVIS21 (50.1% AP) and OVIS (35.5% AP). Furthermore, we find that pseudo-videos transformed from images can train robust models surpassing fully-supervised ones.
****
Unsupervised Video Object Segmentation with Online Adversarial Self-Tuning
Tiankang Su, Huihui Song, Dong Liu, Bo Liu, Qingshan Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 688-698
The existing unsupervised video object segmentation methods depend heavily on the segmentation model trained offline on a labeled training video set, and cannot well generalize to the test videos from a different domain with possible distribution shifts. We propose to perform online fine-tuning on the pre-trained segmentation model to adapt to any ad-hoc videos at the test time. To achieve this, we design an offline semi-supervised adversarial training process, which leverages the unlabeled video frames to improve the model generalizability while aligning the features of the labeled video frames with the features of the unlabeled video frames. With the trained segmentation model, we further conduct an online self-supervised adversarial finetuning, in which a teacher model and a student model are first initialized with the pre-trained segmentation model weights, and the pseudo label produced by the teacher model is used to supervise the student model in an adversarial learning framework. Through online finetuning, the student model is progressively updated according to the emerging patterns in
each test video, which significantly reduces the test-time domain gap. We integrate our offline training and online fine-tuning in a unified framework for unsupervised video object segmentation and dub our method Online Adversarial Self-Tuning (OAST). The experiments show that our method out-performs the state-of-the-arts with significant gains on the popular video object segmentation datasets.
****
Hallucination Improves the Performance of Unsupervised Visual Representation Learning
Jing Wu, Jennifer Hobbs, Naira Hovakimyan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16132-16143
Contrastive learning models based on Siamese structure have demonstrated remarkable performance in self-supervised learning. Such a success of contrastive learning relies on two conditions, including a sufficient number of positive pairs and adequate variations between them. If the conditions are not met, these frameworks will lack semantic contrast and be fragile on overfitting. To address these two issues, we propose Hallucinator that could efficiently generate additional positive samples for further contrast. The Hallucinator creates new data in the feature space, thus introducing nearly negligible computation. Moreover, we reduc

e the mutual information of hallucinated pairs and smooth them through non-linea
r operations. This process helps avoid over-confident contrastive learning model
s during the training and achieves more robust transformation-invariant feature
embeddings. Remarkably, we empirically prove that the proposed Hallucinator gene
ralizes well to various contrastive learning models, including MoCoV1&V2, SimCLR
 and SimSiam. Under the linear classification protocol, a stable accuracy gain i
s achieved, ranging from 0.3% to 3.0% on CIFAR10&100, Tiny ImageNet, STL-10 and
ImageNet. The improvement is also observed in transferring pre-train encoders to
 the downstream tasks, including object detection and segmentation.
****
S3IM: Stochastic Structural SIMilarity and Its Unreasonable Effectiveness for Ne
ural Fields
Zeke Xie, Xindi Yang, Yujie Yang, Qi Sun, Yixiang Jiang, Haoran Wang, Yunfeng Ca
i, Mingming Sun; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 18024-18034
Recently, Neural Radiance Field (NeRF) has shown great success in rendering nove
l-view images of a given scene by learning an implicit representation with only
posed RGB images. NeRF and relevant neural field methods (e.g., neural surface r
epresentation) typically optimize a point-wise loss and make point-wise predicti
ons, where one data point corresponds to one pixel. Unfortunately, this line of
research failed to use the collective supervision of distant pixels, although it
 is known that pixels in an image or scene can provide rich structural informati
on. To the best of our knowledge, we are the first to design a nonlocal multiple
x training paradigm for NeRF and relevant neural field methods via a novel Stoch
astic Structural SIMilarity (S3IM) loss that processes multiple data points as a
 whole set instead of process multiple inputs independently. Our extensive exper
iments demonstrate the unreasonable effectiveness of S3IM in improving NeRF and
neural surface representation for nearly free. The improvements of quality metri
cs can be particularly significant for those relatively difficult tasks: e.g., t
he test MSE loss unexpectedly drops by more than 90% for TensoRF and DVGO over e
ight novel view synthesis tasks; a 198% F-score gain and a 64% Chamfer L1 distan
ce reduction for NeuS over eight surface reconstruction tasks. Moreover, S3IM is
 consistently robust even with sparse inputs, corrupted images, and dynamic scen
es.
****
GlobalMapper: Arbitrary-Shaped Urban Layout Generation
Liu He, Daniel Aliaga; Proceedings of the IEEE/CVF International Conference on C
omputer Vision (ICCV), 2023, pp. 454-464
Modeling and designing urban building layouts is of significant interest in comp
uter vision, computer graphics, and urban applications. A building layout consis
ts of a set of buildings in city blocks defined by a network of roads. We observ
e that building layouts are discrete structures, consisting of multiple rows of
buildings of various shapes, and are amenable to skeletonization for mapping arb
itrary city block shapes to a canonical form. Hence, we propose a fully automati
c approach to building layout generation using a graph attention networks. Our m
ethod generates realistic urban layouts given arbitrary road networks, and enabl
es conditional generation based on learned priors. Our results, including user s
tudy, demonstrate superior performance as compared to prior layout generation ne
tworks, support arbitrary city block and varying building shapes as demonstrated
 by generating layouts for 28 large cities.
****
Membrane Potential Batch Normalization for Spiking Neural Networks
Yufei Guo, Yuhan Zhang, Yuanpei Chen, Weihang Peng, Xiaode Liu, Liwen Zhang, Xuh
ui Huang, Zhe Ma; Proceedings of the IEEE/CVF International Conference on Comput
er Vision (ICCV), 2023, pp. 19420-19430
As one of the energy-efficient alternatives of conventional neural networks (CNN
s), spiking neural networks (SNNs) have gained more and more interest recently.
To train the deep models, some effective batch normalization (BN) techniques are
 proposed in SNNs. All these BNs are suggested to be used after the convolution
layer as usually doing in CNNs. However, the spiking neuron is much more complex

with spatiotemporal dynamics. The regulated data flow after the BN layer will b
e disturbed again by the membrane potential updating operation before the firing
function, i.e., the nonlinear activation. Therefore, we advocate adding another
BN layer before the firing function to normalize the membrane potential again,
called MPBN. To eliminate the induced time cost of MPBN, we also propose a train
ing-inference-decoupled re-parameterization technique to fold the trained MPBN i
nto the firing threshold. With the re-parameterization technique, the MPBN will
not induce any extra time burden in the inference. Furthermore, the MPBN can als
o adopt the element-wised form, while the BN after the convolution layer can onl
y use the channel-wised form. Experimental results show that the proposed MPBN p
erforms well on both popular non-spiking static and neuromorphic datasets.
****
Enhancing Sample Utilization through Sample Adaptive Augmentation in Semi-Superv
ised Learning
Guan Gui, Zhen Zhao, Lei Qi, Luping Zhou, Lei Wang, Yinghuan Shi; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1588
0-15889
In semi-supervised learning, unlabeled samples can be utilized through augmentat
ion and consistency regularization. However, we observed certain samples, even u
ndergoing strong augmentation, are still correctly classified with high confiden
ce, resulting in a loss close to zero. It indicates that these samples have been
already learned well and do not provide any additional optimization benefits to
the model. We refer to these samples as "naive samples". Unfortunately, existin
g SSL models overlook the characteristics of naive samples, and they just apply
the same learning strategy to all samples. To further optimize the SSL model, we
emphasize the importance of giving attention to naive samples and augmenting th
em in a more diverse manner. Sample adaptive augmentation (SAA) is proposed for
this stated purpose and consists of two modules: 1) sample selection module; 2)
sample augmentation module. Specifically, the sample selection module picks out
naive samples based on historical training information at each epoch, then the
naive samples will be augmented in a more diverse manner in the sample augmenta
tion module. Thanks to the extreme ease of implementation of the above modules,
SAA is advantageous for being simple and lightweight. We add SAA on top of FixMa
tch and FlexMatch respectively, and experiments demonstrate SAA can significantl
y improve the models. For example, SAA helped improve the accuracy of FixMatch f
rom 92.50% to 94.76% and that of FlexMatch from 95.01% to 95.31% on CIFAR-10 wit
h 40 labels. The code can be downloaded in supplementary materials.
****
Imitator: Personalized Speech-driven 3D Facial Animation
Balamurugan Thambiraja, Ikhsanul Habibie, Sadegh Aliakbarian, Darren Cosker, Chr
istian Theobalt, Justus Thies; Proceedings of the IEEE/CVF International Confere
nce on Computer Vision (ICCV), 2023, pp. 20621-20631
Speech-driven 3D facial animation has been widely explored, with applications in
gaming, character animation, virtual reality, and telepresence systems. State-o
f-the-art methods deform the face topology of the target actor to sync the input
audio without considering the identity-specific speaking style and facial idios
yncrasies, thus, resulting in unrealistic and inaccurate lip movements. To addre
ss this, we present Imitator, a speech-driven facial expression synthesis method
, which learns identity-specific details from a short input video and produces n
ovel facial expressions matching the identity-specific speaking style and facial
idiosyncrasies of the target actor. Specifically, we train a style-agnostic tra
nsformer on a large facial expression dataset which we use as a prior for audio-
driven facial expressions. We utilize this prior to optimize for identity-specif
ic speaking style based on a short reference video. To train the prior, we intro
duce a novel loss function based on detected bilabial consonants to ensure plaus
ible lip closures and consequently improve the realism of the generated expressi
ons. Through detailed experiments and user studies, we show that our approach im
proves Lip-Sync by 49% and produces expressive facial animations from input audi
o while preserving the actor's speaking style. Project page: https://balamurugan
thambiraja.github.io/Imitator

****
Unified Coarse-to-Fine Alignment for Video-Text Retrieval
Ziyang Wang, Yi-Lin Sung, Feng Cheng, Gedas Bertasius, Mohit Bansal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2816-2827
The canonical approach to video-text retrieval leverages a coarse-grained or fine-grained alignment between visual and textual information. However, retrieving the correct video according to the text query is often challenging as it requires the ability to reason about both high-level (scene) and low-level (object) visual clues and how they relate to the text query. To this end, we propose a Unified Coarse-to-fine Alignment model, dubbed UCOFIA. Specifically, our model captures the cross-modal similarity information at different granularity levels. To alleviate the effect of irrelevant visual clues, we also apply an Interactive Similarity Aggregation module (ISA) to consider the importance of different visual features while aggregating the cross-modal similarity to obtain a similarity score for each granularity. Finally, we apply the Sinkhorn-Knopp algorithm to normalize the similarities of each level before summing them, alleviating over- and under-representation issues at different levels. By jointly considering the cross-modal similarity of different granularity, UCOFIA allows the effective unification of multi-grained alignments. Empirically, UCOFIA outperforms previous state-of-the-art CLIP-based methods on multiple video-text retrieval benchmarks, achieving 2.4%, 1.4% and 1.3% improvements in text-to-video retrieval R@1 on MSR-VTT, Activity-Net, and DiDeMo, respectively. Our code is publicly available at https://github.com/Ziyang412/UCoFiA.
****
Seeing Beyond the Patch: Scale-Adaptive Semantic Segmentation of High-resolution Remote Sensing Imagery based on Reinforcement Learning
Yinhe Liu, Sunan Shi, Junjue Wang, Yanfei Zhong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16868-16878
In remote sensing imagery analysis, patch-based methods have limitations in capturing information beyond the sliding window. This shortcoming poses a significant challenge in processing complex and variable geo-objects, which results in semantic inconsistency in segmentation results. To address this challenge, we propose a dynamic scale perception framework, named GeoAgent, which adaptively captures appropriate scale context information outside the image patch based on the different geo-objects. In GeoAgent, each image patch's states are represented by a global thumbnail and a location mask. The global thumbnail provides context beyond the patch, and the location mask guides the perceived spatial relationships. The scale-selection actions are performed through a Scale Control Agent (SCA). A feature indexing module is proposed to enhance the ability of the agent to distinguish the current image patch's location. The action switches the patch scale and context branch of a dual-branch segmentation network that extracts and fuses the features of multi-scale patches. The GeoAgent adjusts the network parameters to perform the appropriate scale-selection action based on the reward received for the selected scale. The experimental results, using two publicly available datasets and our newly constructed dataset WUSU, demonstrate that GeoAgent outperforms previous segmentation methods, particularly for large-scale mapping applications.
****
Gradient-Regulated Meta-Prompt Learning for Generalizable Vision-Language Models
Juncheng Li, Minghe Gao, Longhui Wei, Siliang Tang, Wenqiao Zhang, Mengze Li, Wei Ji, Qi Tian, Tat-Seng Chua, Yueting Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2551-2562
Prompt tuning, a recently emerging paradigm, enables the powerful vision-language pre-training models to adapt to downstream tasks in a parameter- and data- efficient way, by learning the "soft prompts" to condition frozen pre-training models. Though effective, it is particularly problematic in the few-shot scenario, where prompt tuning performance is sensitive to the initialization and requires a time-consuming process to find a good initialization, thus restricting the fast adaptation ability of the pre-training models. In addition, prompt tuning could

undermine the generalizability of the pre-training models, because the learnable prompt tokens are easy to overfit to the limited training samples. To address these issues, we introduce a novel Gradient-RegulAted Meta-prompt learning (GRAM) framework that jointly meta-learns an efficient soft prompt initialization for better adaptation and a lightweight gradient regulating function for strong cross-domain generalizability in a meta-learning paradigm using only the unlabeled image-text pre-training data. Rather than designing a specific prompt tuning method, our GRAM can be easily incorporated into various prompt tuning methods in a model-agnostic way, and comprehensive experiments show that GRAM brings about consistent improvement for them in several settings (i.e., few-shot learning, cross-domain generalization, cross-dataset generalization, etc.) over 11 datasets. Further, experiments show that GRAM enables the orthogonal methods of textual and visual prompt tuning to work in a mutually-enhanced way, offering better generalizability beyond the uni-modal prompt tuning methods.
****
Zero-Shot Composed Image Retrieval with Textual Inversion
Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, Alberto Del Bimbo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15338-15347
Composed Image Retrieval (CIR) aims to retrieve a target image based on a query composed of a reference image and a relative caption that describes the difference between the two images. The high effort and cost required for labeling datasets for CIR hamper the widespread usage of existing methods, as they rely on supervised learning. In this work, we propose a new task, Zero-Shot CIR (ZS-CIR), that aims to address CIR without requiring a labeled training dataset. Our approach, named zero-Shot composEd imAge Retrieval with textuaL invErsion SEARLE, maps the visual features of the reference image into a pseudo-word token in CLIP token embedding space and integrates it with the relative caption. To support research on ZS-CIR, we introduce an open-domain benchmarking dataset named Composed Image Retrieval on Common Objects in context (CIRCO), which is the first dataset for CIR containing multiple ground-truths for each query. The experiments show that SEARLE exhibits better performance than the baselines on the two main datasets for CIR tasks, FashionIQ and CIRR, and on the proposed CIRCO. The dataset, the code and the model are publicly available at https://github.com/miccunifi/SEARLE.
****
MUter: Machine Unlearning on Adversarially Trained Models
Junxu Liu, Mingsheng Xue, Jian Lou, Xiaoyu Zhang, Li Xiong, Zhan Qin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4892-4902
Machine unlearning is an emerging task of removing the influence of selected training datapoints from a trained model upon data deletion requests, which echoes the widely enforced data regulations mandating the Right to be Forgotten. Many unlearning methods have been proposed recently, achieving significant efficiency gains over the naive baseline of retraining from scratch. However, existing methods focus exclusively on unlearning from standard training models and do not apply to adversarial training models (ATMs) despite their popularity as effective defenses against adversarial examples. During adversarial training, the training data are involved in not only an outer loop for minimizing the training loss, but also an inner loop for generating the adversarial perturbation. Such bi-level optimization greatly complicates the influence measure for the data to be deleted and renders the unlearning more challenging than standard model training with single-level optimization.
This paper proposes a new approach called MUter for unlearning from ATMs. We derive a closed-form unlearning step underpinned by a total Hessian-related data influence measure, while existing methods can mis-capture the data influence associated with the indirect Hessian part. We further alleviate the computational cost by introducing a series of approximations and conversions to avoid the most computationally demanding parts of Hessian inversions. The efficiency and effectiveness of MUter have been validated through experiments on four datasets using b

oth linear and neural network models.
****
WALDO: Future Video Synthesis Using Object Layer Decomposition and Parametric Fl ow Prediction
Guillaume Le Moing, Jean Ponce, Cordelia Schmid; Proceedings of the IEEE/CVF Int ernational Conference on Computer Vision (ICCV), 2023, pp. 23229-23241
This paper presents WALDO (WArping Layer-Decomposed Objects), a novel approach t o the prediction of future video frames from past ones. Individual images are de composed into multiple layers combining object masks and a small set of control points. The layer structure is shared across all frames in each video to build d ense inter-frame connections. Complex scene motions are modeled by combining par ametric geometric transformations associated with individual layers, and video s ynthesis is broken down into discovering the layers associated with past frames, predicting the corresponding transformations for upcoming ones and warping the associated object regions accordingly, and filling in the remaining image parts. Extensive experiments on multiple benchmarks including urban videos (Cityscapes and KITTI) and videos featuring nonrigid motions (UCF-Sports and H3.6M), show t hat our method consistently outperforms the state of the art by a significant ma rgin in every case. Code, pretrained models, and video samples synthesized by ou r approach can be found in the project webpage.
****
ParCNetV2: Oversized Kernel with Enhanced Attention
Ruihan Xu, Haokui Zhang, Wenze Hu, Shiliang Zhang, Xiaoyu Wang; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5752-5 762
Transformers have shown great potential in various computer vision tasks. By bor rowing design concepts from transformers, many studies revolutionized CNNs and s howed remarkable results. This paper falls in this line of studies. Specifically , we propose a new convolutional neural network, ParCNetV2, that extends the res earch line of ParCNetV1 by bridging the gap between CNN and ViT. It introduces t wo key designs: 1) Oversized Convolution (OC) with twice the size of the input, and 2) Bifurcate Gate Unit (BGU) to ensure that the model is input adaptive. Fus ing OC and BGU in a unified CNN, ParCNetV2 is capable of flexibly extracting glo bal features like ViT, while maintaining lower latency and better accuracy. Exte nsive experiments demonstrate the superiority of our method over other convoluti onal neural networks and hybrid models that combine CNNs and transformers. The c ode is publicly available at https://github.com/XuRuihan/ParCNetV2.
****
BiFF: Bi-level Future Fusion with Polyline-based Coordinate for Interactive Traj ectory Prediction
Yiyao Zhu, Di Luan, Shaojie Shen; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2023, pp. 8260-8271
Predicting future trajectories of surrounding agents is essential for safety-cri tical autonomous driving. Most existing work focuses on predicting marginal traj ectories for each agent independently. However, it has rarely been explored in p redicting joint trajectories for interactive agents. In this work, we propose Bi -level Future Fusion (BiFF) to explicitly capture future interactions between in teractive agents. Concretely, BiFF fuses the high-level future intentions follow ed by low-level future behaviors. Then the polyline-based coordinate is specific ally designed for multi-agent prediction to ensure data efficiency, frame robust ness, and prediction accuracy. Experiments show that BiFF achieves state-of-the- art performance on the interactive prediction benchmark of Waymo Open Motion Dat aset.
****
RealGraph: A Multiview Dataset for 4D Real-world Context Graph Generation
Haozhe Lin, Zequn Chen, Jinzhi Zhang, Bing Bai, Yu Wang, Ruqi Huang, Lu Fang; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 23, pp. 3758-3768
In this paper, we propose a brand new scene understanding paradigm called "Conte xt Graph Generation (CGG)", aiming at abstracting holistic semantic information

in the complicated 4D world. The CGG task capitalizes on the calibrated multiview videos of a dynamic scene, and targets at recovering semantic information (coordination, trajectories and relationships) of the presented objects in the form of spatio-temporal context graph in 4D space. We also present a benchmark 4D video dataset "RealGraph", the first dataset tailored for the proposed CGG task. The raw data of RealGraph is composed of calibrated and synchronized multiview videos. We exclusively provide manual annotations including object 2D&3D bounding boxes, category labels and semantic relationships. We also make sure the annotated ID for every single object is temporally and spatially consistent. We propose the first CGG baseline algorithm, Multiview-based Context Graph Generation Network (MCGNet), to empirically investigate the legitimacy of CGG task on RealGraph dataset. We nevertheless reveal the great challenges behind this task and encourage the community to explore beyond our solution.
****
COOL-CHIC: Coordinate-based Low Complexity Hierarchical Image Codec
Théo Ladune, Pierrick Philippe, Félix Henry, Gordon Clare, Thomas Leguay; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13515-13522
We introduce COOL-CHIC, a Coordinate-based Low Complexity Hierarchical Image Codec. It is a learned alternative to autoencoders with 629 parameters and 680 multiplications per decoded pixel. COOL-CHIC offers compression performance close to modern conventional MPEG codecs such as HEVC and is competitive with popular autoencoder-based systems. This method is inspired by Coordinate-based Neural Representations, where an image is represented as a learned function which maps pixel coordinates to RGB values. The parameters of the mapping function are then sent using entropy coding. At the receiver side, the compressed image is obtained by evaluating the mapping function for all pixel coordinates. COOL-CHIC implementation is made open-source.
****
Normalizing Flows for Human Pose Anomaly Detection
Or Hirschorn, Shai Avidan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13545-13554
Video anomaly detection is an ill-posed problem because it relies on many parameters such as appearance, pose, camera angle, background, and more.
We distill the problem to anomaly detection of human pose, thus decreasing the risk of nuisance parameters such as appearance affecting the result. Focusing on pose alone also has the side benefit of reducing bias against distinct minority groups.
Our model works directly on human pose graph sequences and is exceptionally lightweight (1K parameters), capable of running on any machine able to run the pose estimation with negligible additional resources.
We leverage the highly compact pose representation in a normalizing flows framework, which we extend to tackle the unique characteristics of spatio-temporal pose data and show its advantages in this use case.
The algorithm is quite general and can handle training data of only normal examples as well as a supervised setting that consists of labeled normal and abnormal examples.
We report state-of-the-art results on two anomaly detection benchmarks - the unsupervised ShanghaiTech dataset and the recent supervised UBnormal dataset.
Code available at https://github.com/orhir/STG-NF.
****
Reconstructing Groups of People with Hypergraph Relational Reasoning
Buzhen Huang, Jingyi Ju, Zhihao Li, Yangang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14873-14883
Due to the mutual occlusion, severe scale variation, and complex spatial distribution, the current multi-person mesh recovery methods cannot produce accurate absolute body poses and shapes in large-scale crowded scenes. To address the obstacles, we fully exploit crowd features for reconstructing groups of people from a monocular image. A novel hypergraph relational reasoning network is proposed to formulate the complex and high-order relation correlations among individuals an

d groups in the crowd. We first extract compact human features and location information from the original high-resolution image. By conducting the relational reasoning on the extracted individual features, the underlying crowd collectiveness and interaction relationship can provide additional group information for the reconstruction. Finally, the updated individual features and the localization information are used to regress human meshes in camera coordinates. To facilitate the network training, we further build pseudo ground-truth on two crowd datasets, which may also promote future research on pose estimation and human behavior understanding in crowded scenes. The experimental results show that our approach outperforms other baseline methods both in crowded and common scenarios. The code and datasets are publicly available at https://github.com/boycehbz/GroupRec.
****
PivotNet: Vectorized Pivot Learning for End-to-end HD Map Construction
Wenjie Ding, Limeng Qiao, Xi Qiu, Chi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3672-3682
Vectorized high-definition map online construction has garnered considerable attention in the field of autonomous driving research. Most existing approaches model changeable map elements using a fixed number of points, or predict local maps in a two-stage autoregressive manner, which may miss essential details and lead to error accumulation. Towards precise map element learning, we propose a simple yet effective architecture named PivotNet, which adopts unified pivot-based map representations and is formulated as a direct set prediction paradigm. Concretely, we first propose a novel Point-to-Line Mask module to encode both the subordinate and geometrical point-line priors in the network. Then, a well-designed Pivot Dynamic Matching module is proposed to model the topology in dynamic point sequences by introducing the concept of sequence matching. Furthermore, to supervise the position and topology of the vectorized point predictions, we propose a Dynamic Vectorized Sequence loss. Extensive experiments and ablations show that PivotNet is remarkably superior to other SOTAs by 5.9 mAP at least. The code will be available soon.
****
Universal Domain Adaptation via Compressive Attention Matching
Didi Zhu, Yinchuan Li, Junkun Yuan, Zexi Li, Kun Kuang, Chao Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6974-6985
Universal domain adaptation (UniDA) aims to transfer knowledge from the source domain to the target domain without any prior knowledge about the label set. The challenge lies in how to determine whether the target samples belong to common categories. The mainstream methods make judgments based on the sample features, which overemphasizes global information while ignoring the most crucial local objects in the image, resulting in limited accuracy. To address this issue, we propose a Universal Attention Matching (UniAM) framework by exploiting the self-attention mechanism in vision transformer to capture the crucial object information. The proposed framework introduces a novel Compressive Attention Matching (CAM) approach to explore the core information by compressively representing attentions. Furthermore, CAM incorporates a residual-based measurement to determine the sample commonness. By utilizing the measurement, UniAM achieves domain-wise and category-wise Common Feature Alignment (CFA) and Target Class Separation (TCS). Notably, UniAM is the first method utilizing the attention in vision transformer directly to perform classification tasks. Extensive experiments show that UniAM outperforms the current state-of-the-art methods on various benchmark datasets.
****
Contactless Pulse Estimation Leveraging Pseudo Labels and Self-Supervision
Zhihua Li, Lijun Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20588-20597
Remote photoplethysmography (rPPG) is a promising research area involving non-invasive monitoring of vital signs using cameras. While several supervised methods have been proposed, recent research has focused on contrastive-based self-supervised methods. However, these methods often collapse to learning irrelevant periodicities when dealing with interferences such as head motions, facial dynamics,

and video compression. To address this limitation, firstly, we enhance the current self-supervised learning by introducing more reliable and explicit contrastive constraints. Secondly, we propose an innovative learning strategy that seamlessly integrates self-supervised constraints with pseudo-supervisory signals derived from traditional unsupervised methods. This is followed by a co-rectification technique designed to mitigate the adverse effects of noisy pseudo-labels. Experimental results demonstrate the superiority of our methodology over representative models when applied to small, high-quality datasets such as PURE and UBFC-rPPG. Importantly, on large-scale challenging datasets such as VIPL-HR and V4V, our method, with zero annotation cost, not only significantly surpasses prevailing self-supervised techniques but also showcases remarkable alignment with state-of-the-art supervised methods.
****

Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions

Ayaan Haque, Matthew Tancik, Alexei A. Efros, Aleksander Holynski, Angjoo Kanazawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19740-19750

We propose a method for editing NeRF scenes with text-instructions. Given a NeRF of a scene and the collection of images used to reconstruct it, our method uses an image-conditioned diffusion model (InstructPix2Pix) to iteratively edit the input images while optimizing the underlying scene, resulting in an optimized 3D scene that respects the edit instruction. We demonstrate that our proposed method is able to edit large-scale, real-world scenes, and is able to accomplish more realistic, targeted edits than prior work.
****

Point2Mask: Point-supervised Panoptic Segmentation via Optimal Transport

Wentong Li, Yuqian Yuan, Song Wang, Jianke Zhu, Jianshu Li, Jian Liu, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 572-581

Weakly-supervised image segmentation has recently attracted increasing research attentions, aiming to avoid the expensive pixel-wise labeling. In this paper, we present an effective method, namely Point2Mask, to achieve high-quality panoptic prediction using only a single random point annotation per target for training. Specifically, we formulate the panoptic pseudo-mask generation as an Optimal Transport (OT) problem, where each ground-truth (gt) point label and pixel sample are defined as the label supplier and consumer, respectively. The transportation cost is calculated by the introduced task-oriented maps, which focus on the category-wise and instance-wise differences among the various thing and stuff targets. Furthermore, a centroid-based scheme is proposed to set the accurate unit number for each gt point supplier. Hence, the pseudo-mask generation is converted into finding the optimal transport plan at a globally minimal transportation cost, which can be solved via the Sinkhorn-Knopp Iteration. Experimental results on Pascal VOC and COCO demonstrate the promising performance of our proposed Point2Mask approach to point-supervised panoptic segmentation. Source code is available at: https://github.com/LiWentomng/Point2Mask.
****

Multi-Task Learning with Knowledge Distillation for Dense Prediction

Yangyang Xu, Yibo Yang, Lefei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21550-21559

While multi-task learning (MTL) has become an attractive topic, its training usually poses more difficulties than the single-task case. How to successfully apply knowledge distillation into MTL to improve training efficiency and model performance is still a challenging problem. In this paper, we introduce a new knowledge distillation procedure with an alternative match for MTL of dense prediction based on two simple design principles. First, for memory and training efficiency, we use a single strong multi-task model as a teacher during training instead of multiple teachers, as widely adopted in existing studies. Second, we employ a less sensitive Cauchy-Schwarz (CS) divergence instead of the Kullback-Leibler (KL) divergence and propose a CS distillation loss accordingly. With the less sensitive divergence, our knowledge distillation with an alternative match is applie

d for capturing inter-task and intra-task information between the teacher model and the student model of each task, thereby learning more "dark knowledge" for effective distillation. We conducted extensive experiments on dense prediction datasets, including NYUD-v2 and PASCAL-Context, for multiple vision tasks, such as semantic segmentation, human parts segmentation, depth estimation, surface normal estimation, and boundary detection. The results show that our proposed method decidedly improves model performance and the practical inference efficiency.
****

What Does a Platypus Look Like? Generating Customized Prompts for Zero-Shot Image Classification
Sarah Pratt, Ian Covert, Rosanne Liu, Ali Farhadi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15691-15701
Open-vocabulary models are a promising new paradigm for image classification. Unlike traditional classification models, open-vocabulary models classify among any arbitrary set of categories specified with natural language during inference. This natural language, called "prompts", typically consists of a set of hand-written templates (e.g., "a photo of a   ") which are completed with each of the category names. This work introduces a simple method to generate higher accuracy prompts, without relying on any explicit knowledge of the task domain and with far fewer hand-constructed sentences. To achieve this, we combine open-vocabulary models with large language models (LLMs) to create Customized Prompts via Language models (CuPL, pronounced "couple"). In particular, we leverage the knowledge contained in LLMs in order to generate many descriptive sentences that contain important discriminating characteristics of the image categories. This allows the model to place a greater importance on these regions in the image when making predictions. We find that this straightforward and general approach improves accuracy on a range of zero-shot image classification benchmarks, including over one percentage point gain on ImageNet. Finally, this simple baseline requires no additional training and remains completely zero-shot. Code available at https://github.com/sarahpratt/CuPL.
****

Scene as Occupancy
Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, Hongyang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8406-8415
Human driver can easily describe the complex traffic scene by visual system. Such an ability of precise perception is essential for driver's planning. To achieve this, a geometry-aware representation that quantizes the physical 3D scene into structured grid map with semantic labels per cell, termed as 3D Occupancy, would be desirable. Compared to the form of bounding box, a key insight behind occupancy is that it could capture the fine-grained details of critical obstacles in the scene, and thereby facilitate subsequent tasks. Prior or concurrent literature mainly concentrate on a single scene completion task, where we might argue that the potential of this occupancy representation might obsess broader impact. In this paper, we propose OccNet, a multi-view vision centric pipeline with a cascade and temporal voxel decoder to reconstruct 3D occupancy. At the core of OccNet is a general occupancy embedding to represent 3D physical world. Such a descriptor could be applied towards a wide span of driving tasks, including detection, segmentation and planning. To validate the effectiveness of this new representation and our proposed algorithm, we propose OpenOcc, the first dense high-quality 3D occupancy benchmark built on top of nuScenes. Empirical experiments show that there are evident performance gain across multiple tasks, e.g., motion planning could witness a collision rate reduction by 15%-58%, demonstrating the superiority of our method.
****

U-RED: Unsupervised 3D Shape Retrieval and Deformation for Partial Point Clouds
Yan Di, Chenyangguang Zhang, Ruida Zhang, Fabian Manhardt, Yongzhi Su, Jason Rambach, Didier Stricker, Xiangyang Ji, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8884-8895
In this paper, we propose U-RED, an Unsupervised shape REtrieval and Deformation

pipeline that takes an arbitrary object observation as input, typically captured by RGB images or scans, and jointly retrieves and deforms the geometrically similar CAD models from a pre-established database to tightly match the target. Considering existing methods typically fail to handle noisy partial observations, U-RED is designed to address this issue from two aspects. First, since one partial shape may correspond to multiple potential full shapes, the retrieval method must allow such an ambiguous one-to-many relationship. Thereby U-RED learns to project all possible full shapes of a partial target onto the surface of a unit sphere. Then during inference, each sampling on the sphere will yield a feasible retrieval. Second, since real-world partial observations usually contain noticeable noise, a reliable learned metric that measures the similarity between shapes is necessary for stable retrieval. In U-RED, we design a novel point-wise residual-guided metric that allows noise-robust comparison. Extensive experiments on the synthetic datasets PartNet, ComplementMe and the real-world dataset Scan2CAD demonstrate that U-RED surpasses existing state-of-the-art approaches by 47.3%, 16.7% and 31.6% respectively under Chamfer Distance. Codes and trained models will be released soon.
****
RFLA: A Stealthy Reflected Light Adversarial Attack in the Physical World
Donghua Wang, Wen Yao, Tingsong Jiang, Chao Li, Xiaoqian Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4455-4465
Physical adversarial attacks against deep neural networks (DNNs) have recently gained increasing attention. The current mainstream physical attacks use printed adversarial patches or camouflage to alter the appearance of the target object. However, these approaches generate conspicuous adversarial patterns that show poor stealthiness. Another physical deployable attack is the optical attack, featuring stealthiness while exhibiting weakly in the daytime with sunlight. In this paper, we propose a novel Reflected Light Attack (RFLA), featuring effective and stealthy in both the digital and physical world, which is implemented by placing the color transparent plastic sheet and a paper cut of a specific shape in front of the mirror to create different colored geometries on the target object. To achieve these goals, we devise a general framework based on the circle to model the reflected light on the target object. Specifically, we optimize a circle (composed of a coordinate and radius) to carry various geometrical shapes determined by the optimized angle. The fill color of the geometry shape and its corresponding transparency are also optimized. We extensively evaluate the effectiveness of RFLA on different datasets and models. Experiment results suggest that the proposed method achieves over 99% success rate on different datasets and models in the digital world. Additionally, we verify the effectiveness of the proposed method in different physical environments by using sunlight or a flashlight.
****
Nearest Neighbor Guidance for Out-of-Distribution Detection
Jaewoo Park, Yoon Gyo Jung, Andrew Beng Jin Teoh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1686-1695
Detecting out-of-distribution (OOD) samples are crucial for machine learning models deployed in open-world environments. Classifier-based scores are a standard approach for OOD detection due to their fine-grained detection capability. However, these scores often suffer from overconfidence issues, misclassifying OOD samples distant from the in-distribution region. To address this challenge, we propose a method called Nearest Neighbor Guidance (NNGuide) that guides the classifier-based score to respect the boundary geometry of the data manifold. NNGuide reduces the overconfidence of OOD samples while preserving the fine-grained capability of the classifier-based score. We conduct extensive experiments on ImageNet OOD detection benchmarks under diverse settings, including a scenario where the ID data undergoes natural distribution shift. Our results demonstrate that NNGuide provides a significant performance improvement on the base detection scores, achieving state-of-the-art results on both AUROC, FPR95, and AUPR metrics.
****
PatchCT: Aligning Patch Set and Label Set with Conditional Transport for Multi-L

abel Image Classification

Miaoge Li, Dongsheng Wang, Xinyang Liu, Zequn Zeng, Ruiying Lu, Bo Chen, Mingyuan Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15348-15358

Multi-label image classification is a prediction task that aims to identify more than one label from a given image. This paper considers the semantic consistency of the latent space between the visual patch and linguistic label domains and introduces the conditional transport (CT) theory to bridge the acknowledged gap. While recent cross-modal attention-based studies have attempted to align such two representations and achieved impressive performance, they required carefully-designed alignment modules and extra complex operations in the attention computation. We find that by formulating the multi-label classification as a CT problem, we can exploit the interactions between the image and label efficiently by minimizing the bidirectional CT cost. Specifically, after feeding the images and textual labels into the modality-specific encoders, we view each image as a mixture of patch embeddings and a mixture of label embeddings, which capture the local region features and the class prototypes, respectively. CT is then employed to learn and align those two semantic sets by defining the forward and backward navigators. Importantly, the defined navigators in CT distance model the similarities between patches and labels, which provides an interpretable tool to visualize the learned prototypes. Extensive experiments on three public image benchmarks show that the proposed model consistently outperforms the previous methods.

****

VI-Net: Boosting Category-level 6D Object Pose Estimation via Learning Decoupled Rotations on the Spherical Representations

Jiehong Lin, Zewei Wei, Yabin Zhang, Kui Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14001-14011

Rotation estimation of high precision from an RGB-D object observation is a huge challenge in 6D object pose estimation, due to the difficulty of learning in the non-linear space of SO(3). In this paper, we propose a novel rotation estimation network, termed as VI-Net, to make the task easier by decoupling the rotation as the combination of a viewpoint rotation and an in-plane rotation. More specifically, VI-Net bases the feature learning on the sphere with two individual branches for the estimates of two factorized rotations, where a V-Branch is employed to learn the viewpoint rotation via binary classification on the spherical signals, while another I-Branch is used to estimate the in-plane rotation by transforming the signals to view from the zenith direction. To process the spherical signals, a Spherical Feature Pyramid Network is constructed based on a novel design of SPAtial Spherical Convolution (SPA-SConv), which settles the boundary problem of spherical signals via feature padding and realizesviewpoint-equivariant feature extraction by symmetric convolutional operations. We apply the proposed VI-Net to the challenging task of category-level 6D object pose estimation for predicting the poses of unknown objects without available CAD models; experiments on the benchmarking datasets confirm the efficacy of our method, which outperforms the existing ones with a large margin in the regime of high precision.

****

ICD-Face: Intra-class Compactness Distillation for Face Recognition

Zhipeng Yu, Jiaheng Liu, Haoyu Qin, Yichao Wu, Kun Hu, Jiayi Tian, Ding Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21042-21052

Knowledge distillation is an effective model compression method to improve the performance of a lightweight student model by transferring the knowledge of a well-performed teacher model, which has been widely adopted in many computer vision tasks, including face recognition (FR). The current FR distillation methods usually utilize the Feature Consistency Distillation (FCD) (e.g., L2 distance) on the learned embeddings extracted by the teacher and student models. However, after using FCD, we observe that the intra-class similarities of the student model are lower than the intra-class similarities of the teacher model a lot. Therefore, we propose an effective FR distillation method called ICD-Face by introducing intra-class compactness distillation into the existing distillation framework.

Specifically, in ICD-Face, we first propose to calculate the similarity distributions of the teacher and student models, where the feature banks are introduced to construct
sufficient and high-quality positive pairs. Then, we estimate the probability distributions of the teacher and student models and introduce the Similarity Distribution Consistency (SDC) loss to improve the intra-class compactness of
the student model. Extensive experimental results on multiple benchmark datasets demonstrate the effectiveness of our proposed ICD-Face for face recognition.
****
Diffusion-SDF: Conditional Generative Modeling of Signed Distance Functions
Gene Chou, Yuval Bahat, Felix Heide; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2262-2272
Probabilistic diffusion models have achieved state-of-the-art results for image synthesis, inpainting, and text-to-image tasks. However, they are still in the early stages of generating complex 3D shapes. This work proposes Diffusion-SDF, a generative model for shape completion, single-view reconstruction, and reconstruction of real-scanned point clouds. We use neural signed distance functions (SDFs) as our 3D representation to parameterize the geometry of various signals (e.g., point clouds, 2D images) through neural networks. Neural SDFs are implicit functions and diffusing them amounts to learning the reversal of their neural network weights, which we solve using a custom modulation module. Extensive experiments show that our method is capable of both realistic unconditional generation and conditional generation from partial inputs. This work expands the domain of diffusion models from learning 2D, explicit representations, to 3D, implicit representations. Code is released at https://github.com/princeton-computational-imaging/Diffusion-SDF.
****
Open-Vocabulary Object Detection With an Open Corpus
Jiong Wang, Huiming Zhang, Haiwen Hong, Xuan Jin, Yuan He, Hui Xue, Zhou Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6759-6769
Existing open vocabulary object detection (OVD) works expand the object detector toward open categories by replacing the classifier with the category text embeddings and optimizing the region-text alignment on data of the base categories. However, both the class-agnostic proposal generator and the classifier are biased to the seen classes as demonstrated by the gaps of objectness and accuracy assessment between base and novel classes. In this paper, an open corpus, composed of a set of external object concepts and clustered to several centroids, is introduced to improve the generalization ability in the detector. We propose the generalized objectness assessment (GOAT) in the proposal generator based on the visual-text alignment, where the similarities of visual feature to the cluster centroids are summarized as the objectness. This simple heuristic evaluates objectness with concepts in open corpus and is thus generalized to open categories. We further propose category expanding (CE) with open corpus in two training tasks, which enables the detector to perceive more categories in the feature space and get more reasonable optimization direction. For the classification task, we introduce an open corpus classifier by reconstructing original classifier with similar words in text space. For the image-caption alignment task, the open corpus centroids are incorporated to enlarge the negative samples in the contrastive loss. Extensive experiments demonstrate the effectiveness of GOAT and CE, which greatly improve the performance on novel classes and get new state-of-the-art on the OVD benchmarks.
****
Long-range Multimodal Pretraining for Movie Understanding
Dawit Mureja Argaw, Joon-Young Lee, Markus Woodson, In So Kweon, Fabian Caba Heilbron; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13392-13403
Learning computer vision models from (and for) movies has a long-standing history. While great progress has been attained, there is still a need for a pretrained multimodal model that can perform well in the ever-growing set of movie unders

tanding tasks the community has been establishing. In this work, we introduce Long-range Multimodal Pretraining, a strategy, and a model that leverages movie data to train transferable multimodal and cross-modal encoders. Our key idea is to learn from all modalities in a movie by observing and extracting relationships over a long-range. After pretraining, we run ablation studies on the LVU benchmark and validate our modeling choices and the importance of learning from long-range time spans. Our model achieves state-of-the-art on several LVU tasks while being much more data efficient than previous works. Finally, we evaluate our model's transferability by setting a new state-of-the-art in five different benchmarks.
****
MRM: Masked Relation Modeling for Medical Image Pre-Training with Genetics
Qiushi Yang, Wuyang Li, Baopu Li, Yixuan Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21452-21462
Modern deep learning techniques on automatic multimodal medical diagnosis rely on massive expert annotations, which is time-consuming and prohibitive. Recent masked image modeling (MIM)-based pre-training methods have witnessed impressive advances for learning meaningful representations from unlabeled data and transferring to downstream tasks. However, these methods focus on natural images and ignore the specific properties of medical data, yielding unsatisfying generalization performance on downstream medical diagnosis. In this paper, we aim to leverage genetics to boost image pre-training and present a masked relation modeling (MRM) framework. Instead of explicitly masking input data in previous MIM methods leading to loss of disease-related semantics, we design relation masking to mask out token-wise feature relation in both self- and cross-modality levels, which preserves intact semantics within the input and allows the model to learn rich disease-related information. Moreover, to enhance semantic relation modeling, we propose relation matching to align the sample-wise relation between the intact and masked features. The relation matching exploits inter-sample relation by encouraging global constraints in the feature space to render sufficient semantic relation for feature representation. Extensive experiments demonstrate that the proposed framework is simple yet powerful, achieving state-of-the-art transfer performance on various downstream diagnosis tasks.
****
Adverse Weather Removal with Codebook Priors
Tian Ye, Sixiang Chen, Jinbin Bai, Jun Shi, Chenghao Xue, Jingxia Jiang, Junjie Yin, Erkang Chen, Yun Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12653-12664
Despite recent advancements in unified adverse weather removal methods, there remains a significant challenge of achieving realistic fine-grained texture and reliable background reconstruction to mitigate serious distortions.

Inspired by recent advancements in codebook and vector quantization (VQ) techniques, we present a novel Adverse Weather Removal network with Codebook Priors (AWRCP) to address the problem of unified adverse weather removal. AWRCP leverages high-quality codebook priors derived from undistorted images to recover vivid texture details and faithful background structures. However, simply utilizing high-quality features from the codebook does not guarantee good results in terms of fine-grained details and structural fidelity. Therefore, we develop a deformable cross-attention with sparse sampling mechanism for flexible perform feature interaction between degraded features and high-quality features from codebook priors. In order to effectively incorporate high-quality texture features while maintaining the realism of the details generated by codebook priors, we propose a hierarchical texture warping head that gradually fuses hierarchical codebook prior features into high-resolution features at final restoring stage.

With the utilization of the VQ codebook as a feature dictionary of high quality and the proposed designs, AWRCP can largely improve the restored quality of texture details, achieving the state-of-the-art performance across multiple adverse weather removal benchmark.

****
Spectrum-guided Multi-granularity Referring Video Object Segmentation
Bo Miao, Mohammed Bennamoun, Yongsheng Gao, Ajmal Mian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 920-930
Current referring video object segmentation (R-VOS) techniques extract conditional kernels from encoded (low-resolution) vision-language features to segment the decoded high-resolution features. We discovered that this causes significant feature drift, which the segmentation kernels struggle to perceive during the forward computation. This negatively affects the ability of segmentation kernels. To address the drift problem, we propose a Spectrum-guided Multi-granularity (SgMg) approach, which performs direct segmentation on the encoded features and employs visual details to further optimize the masks. In addition, we propose Spectrum-guided Cross-modal Fusion (SCF) to perform intra-frame global interactions in the spectral domain for effective multimodal representation. Finally, we extend SgMg to perform multi-object R-VOS, a new paradigm that enables simultaneous segmentation of multiple referred objects in a video. This not only makes R-VOS faster, but also more practical. Extensive experiments show that SgMg achieves state-of-the-art performance on four video benchmark datasets, outperforming the nearest competitor by 2.8% points on Ref-YouTube-VOS. Our extended SgMg enables multi-object R-VOS, runs about 3 times faster while maintaining satisfactory performance. Code is available at https://github.com/bo-miao/SgMg.
****
Sound Source Localization is All about Cross-Modal Alignment
Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, Joon Son Chung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7777-7787
Humans can easily perceive the direction of sound sources in a visual scene, termed sound source localization. Recent studies on learning-based sound source localization have mainly explored the problem from a localization perspective. However, prior arts and existing benchmarks do not account for a more important aspect of the problem, cross-modal semantic understanding, which is essential for genuine sound source localization. Cross-modal semantic understanding is important in understanding semantically mismatched audio-visual events, e.g., silent objects, or off-screen sounds. To account for this, we propose a cross-modal alignment task as a joint task with sound source localization to better learn the interaction between audio and visual modalities. Thereby, we achieve high localization performance with strong cross-modal semantic understanding. Our method outperforms the state-of-the-art approaches in both sound source localization and cross-modal retrieval. Our work suggests that jointly tackling both tasks is necessary to conquer genuine sound source localization.
****
MAP: Towards Balanced Generalization of IID and OOD through Model-Agnostic Adapters
Min Zhang, Junkun Yuan, Yue He, Wenbin Li, Zhengyu Chen, Kun Kuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11921-11931
Deep learning has achieved tremendous success in recent years, but most of these successes are built on an independent and identically distributed (IID) assumption. This somewhat hinders the application of deep learning to the more challenging out-of-distribution (OOD) scenarios. Although many OOD methods have been proposed to address this problem and have obtained good performance on testing data that is of major shifts with training distributions, interestingly, we experimentally find that these methods achieve excellent OOD performance by making a great sacrifice of the IID performance. We call this finding the IID-OOD dilemma. Clearly, in real-world applications, distribution shifts between training and testing data are often uncertain, where shifts could be minor, and even close to the IID scenario, and thus it is truly important to design a deep model with the balanced generalization ability between IID and OOD. To this end, in this paper, we investigate an intriguing problem of balancing IID and OOD generalizations and propose a novel Model Agnostic adaPters (MAP) method, which is more reliable a

nd effective for distribution-shift-agnostic real-world data. Our key technical contribution is to use auxiliary adapter layers to incorporate the inductive bias of IID into OOD methods. To achieve this goal, we apply a bilevel optimization to explicitly model and optimize the coupling relationship between the OOD model and auxiliary adapter layers. We also theoretically give a first-order approximation to save computational time. Experimental results on six datasets successfully demonstrate that MAP can greatly improve the performance of IID while achieving good OOD performance.
****
Exploring Group Video Captioning with Efficient Relational Approximation
Wang Lin, Tao Jin, Ye Wang, Wenwen Pan, Linjun Li, Xize Cheng, Zhou Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15281-15290
Current video captioning efforts most focus on describing a single video while the need for captioning videos in groups has increased considerably. In this study, we propose a new task, group video captioning, which aims to infer the desired content among a group of target videos and describe it with another group of related reference videos. This task requires the model to effectively summarize the target videos and accurately describe the distinguishing content compared to the reference videos, and it becomes more difficult as the video length increases. To solve this problem, 1) First, we propose an efficient relational approximation (ERA) to identify the shared content among videos while the complexity is linearly related to the number of videos. 2) Then, we introduce a contextual feature refinery with intra-group self-supervision to capture the contextual information and further refine the common properties. 3) In addition, we construct two group video captioning datasets derived from the YouCook2 and the ActivityNet Captions. The experimental results demonstrate the effectiveness of our method on this new task.
****
ADAPT: Efficient Multi-Agent Trajectory Prediction with Adaptation
Görkay Aydemir, Adil Kaan Akan, Fatma Güney; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8295-8305
Forecasting future trajectories of agents in complex traffic scenes requires reliable and efficient predictions for all agents in the scene. However, existing methods for trajectory prediction are either inefficient or sacrifice accuracy. To address this challenge, we propose ADAPT, a novel approach for jointly predicting the trajectories of all agents in the scene with dynamic weight learning. Our approach outperforms state-of-the-art methods in both single-agent and multi-agent settings on the Argoverse and Interaction datasets, with a fraction of their computational overhead. We attribute the improvement in our performance: first, to the adaptive head augmenting the model capacity without increasing the model size; second, to our design choices in the endpoint-conditioned prediction, reinforced by gradient stopping. Our analyses show that ADAPT can focus on each agent with adaptive prediction, allowing for accurate predictions efficiently. https://KUIS-AI.github.io/adapt
****
TaskExpert: Dynamically Assembling Multi-Task Representations with Memorial Mixture-of-Experts
Hanrong Ye, Dan Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21828-21837
Learning discriminative task-specific features simultaneously for multiple distinct tasks is a fundamental problem in multi-task learning. Recent state-of-the-art models consider directly decoding task-specific features from one shared task-generic feature (e.g., feature from a backbone layer), and utilize carefully designed decoders to produce multi-task features. However, as the input feature is fully shared and each task decoder also shares decoding parameters for different input samples, it leads to a static feature decoding process, producing less discriminative task-specific representations. To tackle this limitation, we propose TaskExpert, a novel multi-task mixture-of-experts model that enables learning multiple representative task-generic feature spaces and decoding task-specific

features in a dynamic manner. Specifically, TaskExpert introduces a set of exper
t networks to decompose the backbone feature into several representative task-ge
neric features. Then, the task-specific features are decoded by using dynamic ta
sk-specific gating networks operating on the decomposed task-generic features. F
urthermore, to establish long-range modeling of the task-specific representation
s from different layers of TaskExpert, we design a multi-task feature memory tha
t updates at each layer and acts as an additional feature expert for dynamic tas
k-specific feature decoding. Extensive experiments demonstrate that our TaskExpe
rt clearly outperforms previous best-performing methods on all 9 metrics of two
competitive multi-task learning benchmarks for visual scene understanding (i.e.,
PASCAL-Context and NYUD-v2). Code and models will be made publicly available.
****
Meta OOD Learning For Continuously Adaptive OOD Detection
Xinheng Wu, Jie Lu, Zhen Fang, Guangquan Zhang; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2023, pp. 19353-19364
Out-of-distribution (OOD) detection is crucial to modern deep learning applicati
ons by identifying and alerting about the OOD samples that should not be tested
or used for making predictions. Current OOD detection methods have made signific
ant progress when in-distribution (ID) and OOD samples are drawn from static dis
tributions. However, this can be unrealistic when applied to real-world systems
which often undergo continuous variations and shifts in ID and OOD distributions
over time. Therefore, for an effective application in real-world systems, the d
evelopment of OOD detection methods that can adapt to these dynamic and evolving
distributions is essential. In this paper, we propose a novel and more realisti
c setting called continuously adaptive out-of-distribution (CAOOD) detection whi
ch targets on developing an OOD detection model that enables dynamic and quick a
daptation to a new arriving distribution, with insufficient ID samples during de
ployment time. To address CAOOD, we develop meta OOD learning (MOL) by designing
a learning-to-adapt diagram such that a good initialized OOD detection model is
learned during the training process. In the testing process, MOL ensures OOD de
tection performance over shifting distributions by quickly adapting to new distr
ibutions with a few adaptations. Extensive experiments on several OOD benchmarks
endorse the effectiveness of our method in preserving both ID classification ac
curacy and OOD detection performance on continuously shifting distributions.
****
MAPConNet: Self-supervised 3D Pose Transfer with Mesh and Point Contrastive Lear
ning
Jiaze Sun, Zhixiang Chen, Tae-Kyun Kim; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2023, pp. 14452-14462
3D pose transfer is a challenging generation task that aims to transfer the pose
of a source geometry onto a target geometry with the target identity preserved.
Many prior methods require keypoint annotations to find correspondence between
the source and target. Current pose transfer methods allow end-to-end correspond
ence learning but require the desired final output as ground truth for supervisi
on. Unsupervised methods have been proposed for graph convolutional models but t
hey require ground truth correspondence between the source and target inputs. We
present a novel self-supervised framework for 3D pose transfer which can be tra
ined in unsupervised, semi-supervised, or fully supervised settings without any
correspondence labels. We introduce two contrastive learning constraints in the
latent space: a mesh-level loss for disentangling global patterns including pose
and identity, and a point-level loss for discriminating local semantics. We dem
onstrate quantitatively and qualitatively that our method achieves state-of-the-
art results in supervised 3D pose transfer, with comparable results in unsupervi
sed and semi-supervised settings. Our method is also generalisable to unseen hum
an and animal data with complex topologies.
****
BlendFace: Re-designing Identity Encoders for Face-Swapping
Kaede Shiohara, Xingchao Yang, Takafumi Taketomi; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2023, pp. 7634-7644
The great advancements of generative adversarial networks and face recognition m

odels in computer vision have made it possible to swap identities on images from single sources. Although a lot of studies seems to have proposed almost satisfactory solutions, we notice previous methods still suffer from an identity-attribute entanglement that causes undesired attributes swapping because widely used identity encoders, e.g., ArcFace, have some crucial attribute biases owing to their pretraining on face recognition tasks. To address this issue, we design BlendFace, a novel identity encoder for face-swapping. The key idea behind BlendFace is training face recognition models on blended images whose attributes are replaced with those of another mitigates inter-personal biases such as hairsyles and head shapes. BlendFace feeds disentangled identity features into generators and guides generators properly as an identity loss function. Extensive experiments demonstrate that BlendFace improves the identity-attribute disentanglement in face-swapping models, maintaining a comparable quantitative performance to previous methods.
****

Test-time Personalizable Forecasting of 3D Human Poses
Qiongjie Cui, Huaijiang Sun, Jianfeng Lu, Weiqing Li, Bin Li, Hongwei Yi, Haofan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 274-283
Current motion forecasting approaches typically train a deep end-to-end model from the source domain data, and then apply it directly to target subjects. Despite promising results, they remain non-optimal, due to privacy considerations, the test person and his/her natural properties (e.g., stature, behavioral trait) are typically unseen/absent in training. In this case, the source pre-trained model has a low ability to adapt to these out-of-source characteristics, resulting in an unreliable prediction. To tackle this issue, we propose a novel helper-predictor test-time personalization approach (H/P-TTP), which allows for a generalizable representation of out-of-source subjects to gain more realistic predictions. Concretely, the helper is preceded by explicit and implicit augmenters, where the former yields noisy sequences to improve robustness, while the latter is to generate novel-domain data with an adversarial learning paradigm. Then, the domain-generalizable learning is achieved where the helper can extract cross-subject invariant-knowledge to update the predictor. At test time, given a new person, the predictor is able to be further optimized to empower personalized capabilities to the specific properties. Under several benchmarks, extensive experiments show that with H/P-TTP, the existing predictive models are significantly improved for various unseen subjects.
****

Few-shot Continual Infomax Learning
Ziqi Gu, Chunyan Xu, Jian Yang, Zhen Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19224-19233
Few-shot continual learning is the ability to continually train a neural network from a sequential stream of few-shot data. In this paper, we propose a Few-shot Continual Infomax Learning (FCIL) framework that makes a deep model to continually/incrementally learn new concepts from few labeled samples, relieving the catastrophic forgetting of past knowledge. Specifically, inspired by the theoretical definition of transfer entropy, we introduce a feature embedding infomax to effectively perform the few-shot learning, which can transfer the strong encoding capability of the base network to learn the feature embedding of these novel classes by maximizing the mutual information of different-level feature distributions. Further, considering that the learned knowledge in the human brain is a generalization of actual information and exists in a certain relational structure, we perform continual structure infomax learning to relieve the catastrophic forgetting problem in the continual learning process. The information structure of this learned knowledge can be preserved through maximizing the mutual information across these continual-changing relations of inter-classes. Comprehensive evaluations on CIFAR100, miniImageNet, and CUB200 datasets demonstrate the superiority of our FCIL when compared against state-of-the-art methods on the few-shot continual learning task.
****

A Parse-Then-Place Approach for Generating Graphic Layouts from Textual Descript
ions
Jiawei Lin, Jiaqi Guo, Shizhao Sun, Weijiang Xu, Ting Liu, Jian-Guang Lou, Dongm
ei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Visio
n (ICCV), 2023, pp. 23622-23631
Creating layouts is a fundamental step in graphic design. In this work, we propo
se to use text as the guidance to create graphic layouts, i.e., Text-to-Layout,
aiming to lower the design barriers. Text-to-Layout is a challenging task, becau
se it needs to consider the implicit, combined, and incomplete layout constraint
s from text, each of which has not been studied in previous work. To address thi
s, we present a two-stage approach, named parse-then-place. The approach introdu
ces an intermediate representation (IR) between text and layout to represent div
erse layout constraints. With IR, Text-to-Layout is decomposed into a parse stag
e and a place stage. The parse stage takes a textual description as input and ge
nerates an IR, in which the implicit constraints from the text are transformed i
nto explicit ones. The place stage generates layouts based on the IR. To model c
ombined and incomplete constraints, we use a Transformer-based layout generation
 model and carefully design a way to represent constraints and layouts as sequen
ces. Besides, we adopt the pretrain-then-finetune strategy to boost the performa
nce of the layout generation model with large-scale unlabeled layouts. To evalua
te our approach, we construct two Text-to-Layout datasets and conduct experiment
s on them. Quantitative results, qualitative analysis, and user studies demonstr
ate our approach's effectiveness.
****
DreamBooth3D: Subject-Driven Text-to-3D Generation
Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenh
all, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan Barron, Yuanzhen Li
, Varun Jampani; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 2349-2359
We present DreamBooth3D, an approach to personalize text-to-3D generative models
 from as few as 3-6 casually captured images of a subject. Our approach combines
 recent advances in personalizing text-to-image models (DreamBooth) with text-to
-3D generation (DreamFusion). We find that naively combining these methods fails
 to yield satisfactory subject-specific 3D assets due to personalized text-to-im
age models overfitting to the input viewpoints of the subject. We overcome this
through a 3-stage optimization strategy where we jointly leverage the 3D consist
ency of neural radiance fields together with the personalization capability of t
ext-to-image models. Our method can produce high-quality, subject-specific 3D as
sets with text-driven modifications such as novel poses, colors and attributes t
hat are not seen in any of the input images of the subject.
****
DARTH: Holistic Test-time Adaptation for Multiple Object Tracking
Mattia Segu, Bernt Schiele, Fisher Yu; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2023, pp. 9717-9727
Multiple object tracking (MOT) is a fundamental component of perception systems
for autonomous driving, and its robustness to unseen conditions is a requirement
 to avoid life-critical failures. Despite the urge of safety in driving systems,
 no solution to the MOT adaptation problem to domain shift in test-time conditio
ns has ever been proposed. However, the nature of a MOT system is manifold - req
uiring object detection and instance association - and adapting all its componen
ts is non-trivial. In this paper, we analyze the effect of domain shift on appea
rance-based trackers, and introduce DARTH, a holistic test-time adaptation frame
work for MOT. We propose a detection consistency formulation to adapt object det
ection in a self-supervised fashion, while adapting the instance appearance repr
esentations via our novel patch contrastive loss. We evaluate our method on a va
riety of domain shifts - including sim-to-real, outdoor-to-indoor, indoor-to-out
door - and substantially improve the source model performance on all metrics. Pr
oject page: https://www.vis.xyz/pub/darth.
****
Multi-interactive Feature Learning and a Full-time Multi-modality Benchmark for

Image Fusion and Segmentation
Jinyuan Liu, Zhu Liu, Guanyao Wu, Long Ma, Risheng Liu, Wei Zhong, Zhongxuan Luo, Xin Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8115-8124
Multi-modality image fusion and segmentation play a vital role in autonomous driving and robotic operation. Early efforts focus on boosting the performance for only one task, e.g., fusion or segmentation, making it hard to reach `Best of Both Worlds'. To overcome this issue, in this paper, we propose a Multi-interactive Feature learning architecture for image fusion and segmentation, namely SegMiF, and exploit dual-task correlation to promote the performance of both tasks. The SegMiF is of a cascade structure, containing a fusion sub-network and a commonly used segmentation sub-network. By slickly bridging intermediate features between two components, the knowledge learned from the segmentation task can effectively assist the fusion task. Also, the benefited fusion network supports the segmentation one to perform more pretentiously. Besides, a hierarchical interactive attention block is established to ensure fine-grained mapping of all the vital information between two tasks, so that the modality/semantic features can be fully mutual-interactive. In addition, a dynamic weight factor is introduced to automatically adjust the corresponding weights of each task, which can balance the interactive feature correspondence and break through the limitation of laborious tuning. Furthermore, we construct a smart multi-wave binocular imaging system and collect a full-time multi-modality benchmark with 15 annotated pixel-level categories for image fusion and segmentation. Extensive experiments on several public datasets and our benchmark demonstrate that the proposed method outputs visually appealing fused images and perform averagely 7.66% higher segmentation mIoU in the real-world scene than the state-of-the-art approaches. The source code and benchmark are available at https://github.com/JinyuanLiu-CV/SegMiF.
****
BaRe-ESA: A Riemannian Framework for Unregistered Human Body Shapes
Emmanuel Hartman, Emery Pierson, Martin Bauer, Nicolas Charon, Mohamed Daoudi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14181-14191
We present Basis Restricted Elastic Shape Analysis (BaRe-ESA), a novel Riemannian framework for human body scan representation, interpolation and extrapolation. BaRe-ESA operates directly on unregistered meshes, i.e., without the need to establish prior point to point correspondences or to assume a consistent mesh structure. Our method relies on a latent space representation, which is equipped with a Riemannian (non-Euclidean) metric associated to an invariant higher-order metric on the space of surfaces. Experimental results on the FAUST and DFAUST datasets show that BaRe-ESA brings significant improvements with respect to previous solutions in terms of shape registration, interpolation and extrapolation. The efficiency and strength of our model is further demonstrated in applications such as motion transfer and random generation of body shape and pose.
****
Skip-Plan: Procedure Planning in Instructional Videos via Condensed Action Space Learning
Zhiheng Li, Wenjia Geng, Muheng Li, Lei Chen, Yansong Tang, Jiwen Lu, Jie Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10297-10306
In this paper, we propose Skip-Plan, a condensed action space learning method for procedure planning in instructional videos. Current procedure planning methods all stick to the state-action pair prediction at every timestep and generate actions adjacently. Although it coincides with human intuition, such a methodology consistently struggles with high-dimensional state supervision and error accumulation on action sequences. In this work, we abstract the procedure planning problem as a mathematical chain model. By skipping uncertain nodes and edges in action chains, we transfer long and complex sequence functions into short but reliable ones in two ways. First, we skip all the intermediate state supervision and only focus on action predictions. Second, we decompose relatively long chains into multiple short sub-chains by skipping unreliable intermediate actions. By thi

s means, our model explores all sorts of reliable sub-relations within an action sequence in the condensed action space. Extensive experiments show Skip-Plan achieves state-of-the-art performance on the CrossTask and COIN benchmarks for procedure planning.
****
A Retrospect to Multi-prompt Learning across Vision and Language
Ziliang Chen, Xin Huang, Quanlong Guan, Liang Lin, Weiqi Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22190-22201
The vision community is undergoing the unprecedented progress with the emergence of Vision-Language Pretraining Models (VLMs). Prompt learning plays as the holy grail of accessing VLMs since it enables their fast adaptation to downstream tasks with limited resources. Whereas existing research milling around single-prompt paradigms, rarely investigate the technical potential behind their multi-prompt learning counterparts. This paper aims to provide a principled retrospect for vision-language multi-prompt learning. We extend the recent constant modality gap phenomenon to learnable prompts and then, justify the superiority of vision-language transfer with multi-prompt augmentation, empirically and theoretically. In terms of this observation, we propose an Energy-based Multi-prompt Learning (EMPL) to generate multiple prompt embeddings by drawing instances from an energy-based distribution, which is implicitly defined by VLMs. So our EMPL is not only parameter-efficient but also rigorously lead to the balance between in-domain and out-of-domain open-vocabulary generalization. Comprehensive experiments have been conducted to justify our claims and the excellence of EMPL.
****
Sparse Instance Conditioned Multimodal Trajectory Prediction
Yonghao Dong, Le Wang, Sanping Zhou, Gang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9763-9772
Pedestrian trajectory prediction is critical in many vision tasks but challenging due to the multimodality of the future trajectory. Most existing methods predict multimodal trajectories conditioned by goals (future endpoints) or instances (all future points). However, goal-conditioned methods ignore the intermediate process and instance-conditioned methods ignore the stochasticity of pedestrian motions. In this paper, we propose a simple yet effective Sparse Instance Conditioned Network (SICNet), which gives a balanced solution between goal-conditioned and instance-conditioned methods. Specifically, SICNet learns comprehensive sparse instances, i.e., representative points of the future trajectory, through a mask generated by a long short-term memory encoder and uses the memory mechanism to store and retrieve such sparse instances. Hence SICNet can decode the observed trajectory into the future prediction conditioned on the stored sparse instance. Moreover, we design a memory refinement module that refines the retrieved sparse instances from the memory to reduce memory recall errors. Extensive experiments on ETH-UCY and SDD datasets show that our method outperforms existing state-of-the-art methods. In addition, ablation studies demonstrate the superiority of our method compared with goal-conditioned and instance-conditioned approaches.
****
Label Shift Adapter for Test-Time Adaptation under Covariate and Label Shifts
Sunghyun Park, Seunghan Yang, Jaegul Choo, Sungrack Yun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16421-16431
Test-time adaptation (TTA) aims to adapt a pre-trained model to the target domain in a batch-by-batch manner during inference. While label distributions often exhibit imbalances in real-world scenarios, most previous TTA approaches typically assume that both source and target domain datasets have balanced label distribution. Due to the fact that certain classes appear more frequently in certain domains (e.g., buildings in cities, trees in forests), it is natural that the label distribution shifts as the domain changes. However, we discover that the majority of existing TTA methods fail to address the coexistence of covariate and label shifts. To tackle this challenge, we propose a novel label shift adapter that can be incorporated into existing TTA approaches to deal with label shifts during the TTA process effectively. Specifically, we estimate the label distribution

of the target domain to feed it into the label shift adapter. Subsequently, the label shift adapter produces optimal parameters for the target label distribution. By predicting only the parameters for a part of the pre-trained source model, our approach is computationally efficient and can be easily applied, regardless of the model architectures. Through extensive experiments, we demonstrate that integrating our strategy with TTA approaches leads to substantial performance improvements under the joint presence of label and covariate shifts.
****

NAPA-VQ: Neighborhood-Aware Prototype Augmentation with Vector Quantization for Continual Learning

Tamasha Malepathirana, Damith Senanayake, Saman Halgamuge; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11674-11684

Catastrophic forgetting; the loss of old knowledge upon acquiring new knowledge, is a pitfall faced by deep neural networks in real-world applications. Many prevailing solutions to this problem rely on storing exemplars (previously encountered data), which may not be feasible in applications with memory limitations or privacy constraints. Therefore, the recent focus has been on Non-Exemplar based Class Incremental Learning (NECIL) where a model incrementally learns about new classes without using any past exemplars. However, due to the lack of old data, NECIL methods struggle to discriminate between old and new classes causing their feature representations to overlap. We propose NAPA-VQ: Neighborhood Aware Prototype Augmentation with Vector Quantization, a framework that reduces this class overlap in NECIL. We draw inspiration from Neural Gas to learn the topological relationships in the feature space, identifying the neighboring classes that are most likely to get confused with each other. This neighborhood information is utilized to enforce strong separation between the neighboring classes as well as to generate old class representative prototypes that can better aid in obtaining a discriminative decision boundary between old and new classes. Our comprehensive experiments on CIFAR-100, TinyImageNet, and ImageNet-Subset demonstrate that NAPA-VQ outperforms the State-of-the-art NECIL methods by an average improvement of 5%, 2%, and 4% in accuracy and 10%, 3%, and 9% in forgetting respectively. Our code can be found in https://github.com/TamashaM/NAPA-VQ.git.
****

Dynamic Snake Convolution Based on Topological Geometric Constraints for Tubular Structure Segmentation

Yaolei Qi, Yuting He, Xiaoming Qi, Yuan Zhang, Guanyu Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6070-6079

Accurate segmentation of topological tubular structures, such as blood vessels and roads, is crucial in various fields, ensuring accuracy and efficiency in downstream tasks. However, many factors complicate the task, including thin local structures and variable global morphologies. In this work, we note the specificity of tubular structures and use this knowledge to guide our DSCNet to simultaneously enhance perception in three stages: feature extraction, feature fusion, and loss constraint. First, we propose a dynamic snake convolution to accurately capture the features of tubular structures by adaptively focusing on slender and tortuous local structures. Subsequently, we propose a multi-view feature fusion strategy to complement the attention to features from multiple perspectives during feature fusion, ensuring the retention of important information from different global morphologies. Finally, a continuity constraint loss function, based on persistent homology, is proposed to constrain the topological continuity of the segmentation better. Experiments on 2D and 3D datasets show that our DSCNet provides better accuracy and continuity on the tubular structure segmentation task compared with several methods.
****

Unsupervised Open-Vocabulary Object Localization in Videos

Ke Fan, Zechen Bai, Tianjun Xiao, Dominik Zietlow, Max Horn, Zixu Zhao, Carl-Johann Simon-Gabriel, Mike Zheng Shou, Francesco Locatello, Bernt Schiele, Thomas Brox, Zheng Zhang, Yanwei Fu, Tong He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13747-13755

In this paper, we show that recent advances in video representation learning and

pre-trained vision-language models allow for substantial improvements in self-supervised video object localization. We propose a method that first localizes objects in videos via a slot attention approach and then assigns text to the obtained slots. The latter is achieved by an unsupervised way to read localized semantic information from the pre-trained CLIP model. The resulting video object localization is entirely unsupervised apart from the implicit annotation contained in CLIP, and it is effectively the first unsupervised approach that yields good results on regular video benchmarks.
****
Dataset Quantization
Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17205-17216
State-of-the-art deep neural networks are trained with large amounts (millions or even billions) of data. The expensive computation and memory costs make it difficult to train them on limited hardware resources, especially for recent popular large language models (LLM) and computer vision models (CV). Recent popular dataset distillation methods are thus developed, aiming to reduce the number of training samples via synthesizing small-scale datasets via gradient matching. However, as the gradient calculation is coupled with the specific network architecture, the synthesized dataset is biased and performs poorly when used for training unseen architectures. To address these limitations, we present dataset quantization (DQ), a
new framework to compress large-scale datasets into small subsets which can be used for training any neural network architectures. Extensive experiments demonstrate that DQ is able to generate condensed small datasets for training
unseen network architectures with state-of-the-art compression ratios for lossless model training. To the best of our knowledge, DQ is the first method that can successfully distill large-scale datasets such as ImageNet-1k with a state-of-the-art compression ratio. Notably, with 60% data from ImageNet and 20% data from Alpaca's instruction tuning data, the models can be trained with negligible or no performance drop for both vision tasks (including classification, semantic segmentation, and object detection) as well as language tasks (including instruction tuning tasks such as BBH and DROP).
****
Unsupervised Video Deraining with An Event Camera
Jin Wang, Wenming Weng, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10831-10840
Current unsupervised video deraining methods are inefficient in modeling the intricate spatio-temporal properties of rain, which leads to unsatisfactory results. In this paper, we propose a novel approach by integrating a bio-inspired event camera into the unsupervised video deraining pipeline, which enables us to capture high temporal resolution information and model complex rain characteristics. Specifically, we first design an end-to-end learning-based network consisting of two modules, the asymmetric separation module and the cross-modal fusion module. The two modules are responsible for segregating the features of the rain-background layer, and for positive enhancement and negative suppression from a cross-modal perspective, respectively. Second, to regularize the network training, we elaborately design a cross-modal contrastive learning method that leverages the complementary information from event cameras, exploring the mutual exclusion and similarity of rain-background layers in different domains. This encourages the deraining network to focus on the distinctive characteristics of each layer and learn a more discriminative representation. Moreover, we construct the first real-world dataset comprising rainy videos and events using a hybrid imaging system. Extensive experiments demonstrate the superior performance of our method on both synthetic and real-world datasets.
****
Overcoming Forgetting Catastrophe in Quantization-Aware Training
Ting-An Chen, De-Nian Yang, Ming-Syan Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17358-17367

Quantization is an effective approach for memory cost reduction by compressing networks to lower bits. However, existing quantization processes learned only from the current data tend to suffer from forgetting catastrophe on streaming data, i.e., significant performance decrement on old task data after being trained on new tasks. Therefore, we propose a lifelong quantization process, LifeQuant, to address the problem. We theoretically analyze the forgetting catastrophe from the shift of quantization search space with the change of data tasks. To overcome the forgetting catastrophe, we first minimize the space shift during quantization and propose Proximal Quantization Space Search (ProxQ), for regularizing the search space during quantization to be close to a pre-defined standard space. Afterward, we exploit replay data (a subset of old task data) for retraining in new tasks to alleviate the forgetting problem. However, the limited amount of replay data usually leads to biased quantization performance toward the new tasks. To address the imbalance issue, we design a Balanced Lifelong Learning (BaLL) Loss to reweight (to increase) the influence of replay data in new task learning, by leveraging the class distributions. Experimental results show that LifeQuant achieves outstanding accuracy performance with a low forgetting rate.
****
DIME-FM : DIstilling Multimodal and Efficient Foundation Models
Ximeng Sun, Pengchuan Zhang, Peizhao Zhang, Hardik Shah, Kate Saenko, Xide Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15521-15533
Large Vision-Language Foundation Models (VLFM), such as CLIP, ALIGN and Florence, are trained on large private datasets of image-caption pairs and achieve superior transferability and robustness on downstream tasks, but they are difficult to use in many practical applications due to their large size, high latency and fixed architectures. Unfortunately, recent works show training a small custom VLFM for resource-limited applications is currently very difficult using public and smaller-scale data. In this paper, we introduce a new distillation mechanism (DIME-FM) that allows us to transfer the knowledge contained in large VLFMs to smaller, customized foundation models using a relatively small amount of inexpensive, unpaired images and sentences. We transfer the knowledge from the pre-trained CLIP-ViT-L/14 model to a ViT-B/32 model, with only 40M public images and 28.4M unpaired public sentences. The resulting model "Distill-ViT-B/32" rivals the CLIP-ViT-B/32 model pre-trained on its private WiT dataset (400M image-text pairs): Distill-ViT-B/32 achieves similar results in terms of zero-shot and linear-probing performance on both ImageNet and the ELEVATER (20 image classification tasks) benchmarks. It also displays comparable robustness when evaluated on five datasets with natural distribution shifts from ImageNet.
****
Boosting Single Image Super-Resolution via Partial Channel Shifting
Xiaoming Zhang, Tianrui Li, Xiaole Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13223-13232
Although deep learning has significantly facilitated the progress of single image super-resolution (SISR) in recent years, it still hits bottlenecks to further improve SR performance with the continuous growth of model scale. Therefore, one of the hotspots in the field is to construct efficient SISR models by elevating the effectiveness of feature representation. In this work, we present a straightforward and generic approach for feature enhancement that can effectively promote the performance of SR models, dubbed partial channel shifting (PCS). Specifically, it is inspired by the temporal shifting in video understanding and displaces part of the channels along the spatial dimensions, thus allowing the effective receptive field to be amplified and the feature diversity to be augmented at almost zero cost. Also, it can be assembled into off-the-shelf models as a plug-and-play component for performance boosting without extra network parameters and computational overhead. However, regulating the features with PCS encounters some issues, like shifting directions and amplitudes, proportions, and patterns of shifted channels, etc. We impose some technical constraints on the issues to simplify the general channel shifting. Extensive and throughout experiments illustrate that the PCS indeed enlarges the effective receptive field, augments the fea

ture diversity for efficiently enhancing SR recovery, and can endow obvious performance gains to existing models.
****
Learning to Upsample by Learning to Sample
Wenze Liu, Hao Lu, Hongtao Fu, Zhiguo Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6027-6037
We present DySample, an ultra-lightweight and effective dynamic upsampler. While impressive performance gains have been witnessed from recent kernel-based dynamic upsamplers such as CARAFE, FADE, and SAPA, they introduce much workload, mostly due to the time-consuming dynamic convolution and the additional sub-network used to generate dynamic kernels. Further, the need for high-res feature guidance of FADE and SAPA somehow limits their application scenarios. To address these concerns, we bypass dynamic convolution and formulate upsampling from the perspective of point sampling, which is more resource-efficient and can be easily implemented with the standard built-in function in PyTorch. We first showcase a naive design, and then demonstrate how to strengthen its upsampling behavior step by step towards our new upsampler, DySample. Compared with former kernel-based dynamic upsamplers, DySample requires no customized CUDA package and has much fewer parameters, FLOPs, GPU memory, and latency. Besides the light-weight characteristics, DySample outperforms other upsamplers across five dense prediction tasks, including semantic segmentation, object detection, instance segmentation, panoptic segmentation, and monocular depth estimation. Code is available at https://github.com/tiny-smart/dysample.
****
LayoutDiffusion: Improving Graphic Layout Generation by Discrete Diffusion Probabilistic Models
Junyi Zhang, Jiaqi Guo, Shizhao Sun, Jian-Guang Lou, Dongmei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7226-7236
Creating graphic layouts is a fundamental step in graphic designs. In this work, we present a novel generative model named LayoutDiffusion for automatic layout generation. As layout is typically represented as a sequence of discrete tokens, LayoutDiffusion models layout generation as a discrete denoising diffusion process. It learns to reverse a mild forward process, in which layouts become increasingly chaotic with the growth of forward steps and layouts in the neighboring steps do not differ too much. Designing such a mild forward process is however very challenging as layout has both categorical attributes and ordinal attributes. To tackle the challenge, we summarize three critical factors for achieving a mild forward process for the layout, i.e., legality, coordinate proximity and type disruption. Based on the factors, we propose a block-wise transition matrix coupled with a piece-wise linear noise schedule. Experiments on RICO and PubLayNet datasets show that LayoutDiffusion outperforms state-of-the-art approaches significantly. Moreover, it enables two conditional layout generation tasks in a plug-and-play manner without re-training and achieves better performance than existing methods. Project page: https://layoutdiffusion.github.io.
****
Efficiently Robustify Pre-Trained Models
Nishant Jain, Harkirat Behl, Yogesh Singh Rawat, Vibhav Vineet; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5505-5515
A recent trend in deep learning algorithms has been towards training large scale models, having high parameter count and trained on big dataset. However, robustness of such large scale models towards real-world settings is still a less-explored topic. In this work, we first benchmark the performance of these models under different perturbations and datasets thereby representing real-world shifts, and highlight their degrading performance under these shifts. We then discuss on how complete model fine-tuning based existing robustification schemes might not be a scalable option given very large scale networks and can also lead them to forget some of the desired characterstics. Finally, we propose a simple and cost-effective method to solve this problem, inspired by knowledge transfer literatu

re. It involves robustifying smaller models, at a lower computation cost, and then use them as teachers to tune a fraction of these large scale networks, reducing the overall computational overhead. We evaluate our proposed method under various vision perturbations including ImageNet-C,R,S,A datasets and also for transfer learning, zero-shot evaluation setups on different datasets. Benchmark results show that our method is able to induce robustness to these large scale models efficiently, requiring significantly lower time and also preserves the transfer learning, zero-shot properties of the original model which none of the existing methods are able to achieve.
****
Efficient Video Prediction via Sparsely Conditioned Flow Matching
Aram Davtyan, Sepehr Sameni, Paolo Favaro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23263-23274
We introduce a novel generative model for video prediction based on latent flow matching, an efficient alternative to diffusion-based models. In contrast to prior work, we keep the high costs of modeling the past during training and inference at bay by conditioning only on a small random set of past frames at each integration step of the image generation process. Moreover, to enable the generation of high-resolution videos and to speed up the training, we work in the latent space of a pretrained VQGAN. Finally, we propose to approximate the initial condition of the flow ODE with the previous noisy frame. This allows to reduce the number of integration steps and hence, speed up the sampling at inference time. We call our model Random frame conditioned flow Integration for VidEo pRediction, or, in short, RIVER. We show that RIVER achieves superior or on par performance compared to prior work on common video prediction benchmarks, while requiring an order of magnitude fewer computational resources. Project website: https://araachie.github.io/river.
****
Surface Normal Clustering for Implicit Representation of Manhattan Scenes
Nikola Popovic, Danda Pani Paudel, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17860-17870
Novel view synthesis and 3D modeling using implicit neural field representation are shown to be very effective for calibrated multi-view cameras. Such representations are known to benefit from additional geometric and semantic supervision. Most existing methods that exploit additional supervision require dense pixel-wise labels or localized scene priors. These methods cannot benefit from high-level vague scene priors provided in terms of scenes' descriptions. In this work, we aim to leverage the geometric prior of Manhattan scenes to improve the implicit neural radiance field representations. More precisely, we assume that only the knowledge of the indoor scene (under investigation) being Manhattan is known -- with no additional information whatsoever -- with an unknown Manhattan coordinate frame. Such high-level prior is used to self-supervise the surface normals derived explicitly in the implicit neural fields. Our modeling allows us to cluster the derived normals and exploit their orthogonality constraints for self-supervision. Our exhaustive experiments on datasets of diverse indoor scenes demonstrate the significant benefit of the proposed method over the established baselines. The source code will be available at https://github.com/nikola3794/normal-clustering-nerf.
****
Distracting Downpour: Adversarial Weather Attacks for Motion Estimation
Jenny Schmalfuss, Lukas Mehl, Andrés Bruhn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10106-10116
Current adversarial attacks on motion estimation, or optical flow, optimize small per-pixel perturbations, which are unlikely to appear in the real world. In contrast, adverse weather conditions constitute a much more realistic threat scenario. Hence, in this work, we present a novel attack on motion estimation that exploits adversarially optimized particles to mimic weather effects like snowflakes, rain streaks or fog clouds. At the core of our attack framework is a differentiable particle rendering system that integrates particles (i) consistently over multiple time steps (ii) into the 3D space (iii) with a photo-realistic appeara

nce. Through optimization, we obtain adversarial weather that significantly impa
cts the motion estimation. Surprisingly, methods that previously showed good rob
ustness towards small per-pixel perturbations are particularly vulnerable to adv
ersarial weather. At the same time, augmenting the training with non-optimized w
eather increases a method's robustness towards weather effects and improves gene
ralizability at almost no additional cost. Our code is available at https://gith
ub.com/cv-stuttgart/DistractingDownpour.
****
Adaptive Similarity Bootstrapping for Self-Distillation Based Representation Lea
rning
Tim Lebailly, Thomas Stegmüller, Behzad Bozorgtabar, Jean-Philippe Thiran, Tinne
 Tuytelaars; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2023, pp. 16505-16514
Most self-supervised methods for representation learning leverage a cross-view c
onsistency objective i.e., they maximize the representation similarity of a give
n image's augmented views. Recent work NNCLR goes beyond the cross-view paradigm
 and uses positive pairs from different images obtained via nearest neighbor boo
tstrapping in a contrastive setting. We empirically show that as opposed to the
contrastive learning setting which relies on negative samples, incorporating nea
rest neighbor bootstrapping in a self-distillation scheme can lead to a performa
nce drop or even collapse. We scrutinize the reason for this unexpected behavior
 and provide a solution. We propose to adaptively bootstrap neighbors based on t
he estimated quality of the latent space. We report consistent improvements comp
ared to the naive bootstrapping approach and the original baselines. Our approac
h leads to performance improvements for various self-distillation method/backbon
e combinations and standard downstream tasks. Our code is publicly available at
https://github.com/tileb1/AdaSim.
****
Generalized Differentiable RANSAC
Tong Wei, Yash Patel, Alexander Shekhovtsov, Jiri Matas, Daniel Barath; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 17649-17660
We propose -RANSAC, a generalized differentiable RANSAC that allows learning the
 entire randomized robust estimation pipeline. The proposed approach enables the
 use of relaxation techniques for estimating the gradients in the sampling distr
ibution, which are then propagated through a differentiable solver. The trainabl
e quality function marginalizes over the scores from all the models estimated wi
thin -RANSAC to guide the network learning accurate and useful inlier probabilit
ies or to train feature detection and matching networks. Our method directly max
imizes the probability of drawing a good hypothesis, allowing us to learn better
 sampling distributions. We test -RANSAC on various real-world scenarios on fund
amental and essential matrix estimation, and 3D point cloud registration, outdoo
rs and indoors, with handcrafted and learning-based features. It is superior to
the state-of-the-art in terms of accuracy while running at a similar speed to it
s less accurate alternatives. The code and trained models are available at https
://github.com/weitong8591/differentiable_ransac.
****
Unfolding Framework with Prior of Convolution-Transformer Mixture and Uncertaint
y Estimation for Video Snapshot Compressive Imaging
Siming Zheng, Xin Yuan; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2023, pp. 12738-12749
We consider the problem of video snapshot compressive imaging (SCI), where seque
ntial high-speed frames are modulated by different masks and captured by a singl
e measurement. The underlying principle of reconstructing multi-frame images fro
m only one single measurement is to solve an ill-posed problem. By combining opt
imization algorithms and neural networks, deep unfolding networks (DUNs) score t
remendous achievements in solving inverse problems. In this paper, our proposed
model is under the DUN framework and we propose a 3D Convolution-Transformer Mix
ture (CTM) module with a 3D efficient and scalable attention model plugged in, w
hich helps fully learn the correlation between temporal and spatial dimensions b

y virtue of
 Transformer. To our best knowledge, this is the first time that Transformer is employed to video SCI reconstruction. Besides, to further investigate the high-frequency information during the reconstruction process which are neglected in previous studies, we introduce variance estimation characterizing the uncertainty on a pixel-by-pixel basis. Extensive experimental results demonstrate that our proposed method achieves state-of-the-art (SOTA) results. Code can be found on https://github.com/zsm1211/CTM-SCI.
****
Non-Semantics Suppressed Mask Learning for Unsupervised Video Semantic Compression
Yuan Tian, Guo Lu, Guangtao Zhai, Zhiyong Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13610-13622
Most video compression methods aim to improve the decoded video visual quality, instead of particularly guaranteeing the semantic-completeness, which deteriorates downstream video analysis tasks, e.g., action recognition. In this paper, we focus on a novel unsupervised video semantic compression problem, where video semantics is compressed in a downstream task-agnostic manner. To tackle this problem, we first propose a Semantic-Mining-then-Compensation (SMC) framework to enhance the plain video codec with powerful semantic coding capability. Then, we optimize the framework with only unlabeled video data, by masking out a proportion of the compressed video and reconstructing the masked regions of the original video, which is inspired by recent masked image modeling (MIM) methods. Although the MIM scheme learns generalizable semantic features, its inner generative learning paradigm may also facilitate the coding framework memorizing non-semantic information with extra bitcosts. To suppress this deficiency, we explicitly decrease the non-semantic information entropy of the decoded video features, by formulating it as a parametrized Gaussian Mixture Model conditioned on the mined video semantics. Comprehensive experimental results demonstrate the proposed approach shows remarkable superiority over previous traditional, learnable and perceptual-quality-oriented video codecs, on three video analysis tasks and seven datasets.
****
ResQ: Residual Quantization for Video Perception
Davide Abati, Haitam Ben Yahia, Markus Nagel, Amirhossein Habibian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17119-17129
This paper accelerates video perception, such as semantic segmentation and human pose estimation, by levering cross-frame redundancies. Unlike the existing approaches, which avoid redundant computations by warping the past features using optical-flow or by performing sparse convolutions on frame differences, we approach the problem from a new perspective: low-bit quantization. We observe that residuals, as the difference in network activations between two neighboring frames, exhibit properties that make them highly quantizable. Based on this observation, we propose a novel quantization scheme for video networks coined as Residual Quantization. ResQ extends the standard, frame-by-frame, quantization scheme by incorporating temporal dependencies that lead to better performance in terms of accuracy vs. bit-width. Furthermore, we extend our model to dynamically adjust the bit-width proportional to the amount of changes in the video. We demonstrate the superiority of our model, against the standard quantization and existing efficient video perception models, using various architectures on semantic segmentation and human pose estimation benchmarks.
****
Inverse Compositional Learning for Weakly-supervised Relation Grounding
Huan Li, Ping Wei, Zeyu Ma, Nanning Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15477-15487
Video relation grounding (VRG) is a significant and challenging problem in the domains of cross-modal learning and video understanding. In this study, we introduce a novel approach called inverse compositional learning (ICL) for weakly-supervised video relation grounding. Our approach represents relations at both the h

olistic and partial levels, formulating VRG as a joint optimization problem that encompasses reasoning at both levels.
For holistic-level reasoning, we propose an inverse attention mechanism and a compositional encoder to generate compositional relevance features. Additionally, we introduce an inverse loss to evaluate and learn the relevance between visual features and relation features.
At the partial-level reasoning, we introduce a grounding by classification scheme. By leveraging the learned holistic-level features and partial-level features, we train the entire model in an end-to-end manner.
We conduct evaluations on two challenging datasets and demonstrate the substantial superiority of our proposed method over state-of-the-art methods. Extensive ablation studies confirm the effectiveness of our approach.
****
XMem++: Production-level Video Segmentation From Few Annotated Frames
Maksym Bekuzarov, Ariana Bermudez, Joon-Young Lee, Hao Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 635-644
Despite advancements in user-guided video segmentation, extracting complex objects consistently for highly complex scenes is still a labor-intensive task, especially for production. It is not uncommon that a majority of frames need to be annotated. We introduce a novel semi-supervised video object segmentation (SSVOS) model, XMem++, that improves existing memory-based models, with a permanent memory module. Most existing methods focus on single frame annotations, while our approach can effectively handle multiple user-selected frames with varying appearances of the same object or region. Our method can extract highly consistent results while keeping the required number of frame annotations low. We further introduce an iterative and attention-based frame suggestion mechanism, which computes the next best frame for annotation. Our method is real-time and does not require retraining after each user input. We also introduce a new dataset, PUMaVOS, which covers new challenging use cases not found in previous benchmarks. We demonstrate SOTA performance on challenging (partial and multi-class) segmentation scenarios as well as long videos, while ensuring significantly fewer frame annotations than any existing method. Project page: https://max810.github.io/xmem2-project-page/
****
MHCN: A Hyperbolic Neural Network Model for Multi-view Hierarchical Clustering
Fangfei Lin, Bing Bai, Yiwen Guo, Hao Chen, Yazhou Ren, Zenglin Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16525-16535
Multi-view hierarchical clustering (MCHC) plays a pivotal role in comprehending the structures within multi-view data, which hinges on the skillful interaction between hierarchical feature learning and comprehensive representation learning across multiple views. However, existing methods often overlook this interplay due to the simple heuristic agglomerative strategies or the decoupling of multi-view representation learning and hierarchical modeling, thus leading to insufficient representation learning. To address these issues, this paper proposes a novel Multi-view Hierarchical Clustering Network (MHCN) model by performing simultaneous multi-view learning and hierarchy modeling. Specifically, to uncover efficient tree-like structures among all views, we derive multiple hyperbolic autoencoders with latent space mapped onto the Poincare ball. Then, the corresponding hyperbolic embeddings are further regularized to achieve the multi-view representation learning principles for both view-common and view-private information, and to ensure hyperbolic uniformity with a well-balanced hierarchy for better interpretability. Extensive experiments on real-world and synthetic multi-view datasets have demonstrated that our method can achieve state-of-the-art hierarchical clustering performance, and empower the clustering results with good interpretability.
****
End-to-End Diffusion Latent Optimization Improves Classifier Guidance
Bram Wallace, Akash Gokul, Stefano Ermon, Nikhil Naik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7280-7290

Classifier guidance---using the gradients of an image classifier to steer the generations of a diffusion model---has the potential to dramatically expand the creative control over image generation and editing. However, currently classifier guidance requires either training new noise-aware models to obtain accurate gradients or using a one-step denoising approximation of the final generation, which leads to misaligned gradients and sub-optimal control. We highlight this approximation's shortcomings and propose a novel guidance method: Direct Optimization of Diffusion Latents (DOODL), which enables plug-and-play guidance by optimizing diffusion latents w.r.t. the gradients of a pre-trained classifier on the true generated pixels, using an invertible diffusion process to achieve memory-efficient backpropagation. Showcasing the potential of more precise guidance, DOODL outperforms one-step classifier guidance on computational and human evaluation metrics across different forms of guidance: using CLIP guidance to improve generations of complex prompts from DrawBench, using fine-grained visual classifiers to expand the vocabulary of Stable Diffusion, enabling image-conditioned generation with a CLIP visual encoder, and improving image aesthetics using an aesthetic scoring network.
****
FineRecon: Depth-aware Feed-forward Network for Detailed 3D Reconstruction
Noah Stier, Anurag Ranjan, Alex Colburn, Yajie Yan, Liang Yang, Fangchang Ma, Baptiste Angles; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18423-18432
Recent works on 3D reconstruction from posed images have demonstrated that direct inference of scene-level 3D geometry without test-time optimization is feasible using deep neural networks, showing remarkable promise and high efficiency. However, the reconstructed geometry, typically represented as a 3D truncated signed distance function (TSDF), is often coarse without fine geometric details. To address this problem, we propose three effective solutions for improving the fidelity of inference-based 3D reconstructions. We first present a resolution-agnostic TSDF supervision strategy to provide the network with a more accurate learning signal during training, avoiding the pitfalls of TSDF interpolation seen in previous work. We then introduce a depth guidance strategy using multi-view depth estimates to enhance the scene representation and recover more accurate surfaces. Finally, we develop a novel architecture for the final layers of the network, conditioning the output TSDF prediction on high-resolution image features in addition to coarse voxel features, enabling sharper reconstruction of fine details. Our method, FineRecon, produces smooth and highly accurate reconstructions, showing significant improvements across multiple depth and 3D reconstruction metrics.
****
Navigating to Objects Specified by Images
Jacob Krantz, Theophile Gervet, Karmesh Yadav, Austin Wang, Chris Paxton, Roozbeh Mottaghi, Dhruv Batra, Jitendra Malik, Stefan Lee, Devendra Singh Chaplot; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10916-10925
Images are a convenient way to specify which particular object instance an embodied agent should navigate to. Solving this task requires semantic visual reasoning and exploration of unknown environments. We present a system that can perform this task in both simulation and the real world. Our modular method solves sub-tasks of exploration, goal instance re-identification, goal localization, and local navigation. We re-identify the goal instance in egocentric vision using feature-matching and localize the goal instance by projecting matched features to a map. Each sub-task is solved using off-the-shelf components requiring zero fine-tuning. On the HM3D InstanceImageNav benchmark, this system outperforms a baseline end-to-end RL policy 7x and outperforms a state-of-the-art ImageNav model 2.3x (56% vs. 25% success). We deploy this system to a mobile robot platform and demonstrate effective performance in the real world, achieving an 88% success rate across a home and an office environment.
****
TRM-UAP: Enhancing the Transferability of Data-Free Universal Adversarial Pertur

bation via Truncated Ratio Maximization
Yiran Liu, Xin Feng, Yunlong Wang, Wu Yang, Di Ming; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4762-4771
Aiming at crafting a single universal adversarial perturbation (UAP) to fool CNN models for various data samples, universal attack enables a more efficient and accurate evaluation for the robustness of CNN models. Early universal attacks craft UAPs depending on data priors. For more practical applications, the data-free universal attacks that make UAPs from random noises have aroused much attention recently. However, existing data-free UAP methods perturb all the CNN feature layers equally via the maximization of the CNN activation, leading to poor transferability. In this paper, we propose a novel data-free universal attack without depending on any real data samples through truncated ratio maximization, which we term as TRM-UAP. Specifically, different from the maximization of the positive activation in convolution layers, we propose to optimize the UAP generation from the ratio of positive and negative activations. To further enhance the transferability of universal attack, TRM-UAP not only performs the ratio maximization merely on low-level generic features via the truncation strategy, but also incorporates a curriculum optimization algorithm that can effectively learn the diversity of artificial images. Extensive experiments on the ImageNet dataset verify that TRM-UAP achieves a state-of-the-art average fooling rate and excellent transferability on different CNN models as compared to other data-free UAP methods. Code is available at https://github.com/RandolphCarter0/TRMUAP.
****
LATR: 3D Lane Detection from Monocular Images with Transformer
Yueru Luo, Chaoda Zheng, Xu Yan, Tang Kun, Chao Zheng, Shuguang Cui, Zhen Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7941-7952
3D lane detection from monocular images is a fundamental yet challenging task in autonomous driving. Recent advances primarily rely on structural 3D surrogates (e.g., bird's eye view) built from front-view image features and camera parameters. However, the depth ambiguity in monocular images inevitably causes misalignment between the constructed surrogate feature map and the original image, posing a great challenge for accurate lane detection. To address the above issue, we present a novel LATR model, an end-to-end 3D lane detector that uses 3D-aware front-view features without transformed view representation. Specifically, LATR detects 3D lanes via cross-attention based on query and key-value pairs, constructed using our lane-aware query generator and dynamic 3D ground positional embedding. On the one hand, each query is generated based on 2D lane-aware features and adopts a hybrid embedding to enhance the lane information. On the other hand, 3D space information is injected as positional embedding from an iteratively-updated 3D ground plane. LATR outperforms previous state-of-the-art methods on both synthetic Apollo and realistic OpenLane, ONCE-3DLanes datasets by large margins (e.g., 11.4 gain in terms of F1 score on OpenLane). Code will be released at https://github.com/JMoonr/LATR.
****
Scratching Visual Transformer's Back with Uniform Attention
Nam Hyeon-Woo, Kim Yu-Ji, Byeongho Heo, Dongyoon Han, Seong Joon Oh, Tae-Hyun Oh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5807-5818
The favorable performance of Vision Transformers (ViTs) is often attributed to the multi-head self-attention (MSA), which enables global interactions at each layer of
a ViT model. Previous works acknowledge the property of long-range dependency for the effectiveness in MSA. In this work, we study the role of MSA in terms of the different axis, density. Our preliminary analyses suggest that the spatial interactions of learned attention maps are close to dense interactions rather than sparse ones. This is a curious phenomenon because dense attention maps are harder for the model to learn due to softmax. We interpret this opposite behavior against softmax as a strong preference for the ViT models to include dense interaction. We thus manually insert the dense uniform attention to each layer of the

ViT models to supply the much-needed dense interactions. We call this method Context Broadcasting, CB. Our study demonstrates the inclusion of CB takes the role of dense attention, and thereby reduces the degree of density in the original attention maps by complying softmax in MSA. We also show that, with negligible costs of CB (1 line in your model code and no additional parameters), both the capacity and generalizability of the ViT models are increased.
****

Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation
Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7623-7633
To replicate the success of text-to-image (T2I) generation, recent works employ large-scale video datasets to train a text-to-video (T2V) generator. Despite their promising results, such paradigm is computationally expensive. In this work, we propose a new T2V generation setting--One-Shot Video Tuning, where only one text-video pair is presented. Our model is built on state-of-the-art T2I diffusion models pre-trained on massive image data. We make two key observations: 1) T2I models can generate still images that represent verb terms; 2) extending T2I models to generate multiple images concurrently exhibits surprisingly good content consistency. To further learn continuous motion, we introduce Tune-A-Video, which involves a tailored spatio-temporal attention mechanism and an efficient one-shot tuning strategy. At inference, we employ DDIM inversion to provide structure guidance for sampling. Extensive qualitative and numerical experiments demonstrate the remarkable ability of our method across various applications.
****

Anchor-Intermediate Detector: Decoupling and Coupling Bounding Boxes for Accurate Object Detection
Yilong Lv, Min Li, Yujie He, Shaopeng Li, Zhuzhen He, Aitao Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6275-6284
Anchor-based detectors have been continuously developed for object detection. However, the individual anchor box makes it difficult to predict the boundary's offset accurately. Instead of taking each bounding box as a closed individual, we consider using multiple boxes together to get prediction boxes. To this end, this paper proposes the Box Decouple-Couple(BDC) strategy in the inference, which no longer discards the overlapping boxes, but decouples the corner points of these boxes. Then, according to each corner's score, we couple the corner points to select the most accurate corner pairs. To meet the BDC strategy, a simple but novel model is designed named the Anchor-Intermediate Detector(AID), which contains two head networks, i.e., an anchor-based head and an anchor-free Corner-aware head. The corner-aware head is able to score the corners of each bounding box to facilitate the coupling between corner points. Extensive experiments on MS COCO show that the proposed anchor-intermediate detector respectively outperforms their baseline RetinaNet and GFL method by 2.4 and 1.2 AP on the MS COCO test-dev dataset without any bells and whistles.
****

Environment-Invariant Curriculum Relation Learning for Fine-Grained Scene Graph Generation
Yukuan Min, Aming Wu, Cheng Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13296-13307
The scene graph generation (SGG) task is designed to identify the predicates based on the subject-object pairs. However, existing datasets generally include two imbalance cases: one is the class imbalance from the predicted predicates and another is the context imbalance from the given subject-object pairs, which presents significant challenges for SGG. Most existing methods focus on the imbalance of the predicted predicate while ignoring the imbalance of the subject-object pairs, which could not achieve satisfactory results. To address the two imbalance cases, we propose a novel Environment Invariant Curriculum Relation learning (EICR) method, which can be applied in a plug-and-play fashion to existing SGG met

hods. Concretely, to remove the imbalance of the subject-object pairs, we first construct different distribution environments for the subject-object pairs and learn a model invariant to the environment changes. Then, we construct a class-balanced curriculum learning strategy to balance the different environments to remove the predicate imbalance. Comprehensive experiments conducted on VG and GQA datasets demonstrate that our EICR framework can be taken as a general strategy for various SGG models, and achieve significant improvements.
****
Extensible and Efficient Proxy for Neural Architecture Search
Yuhong Li, Jiajie Li, Cong Hao, Pan Li, Jinjun Xiong, Deming Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6199-6210
Efficient or near-zero-cost proxies were proposed recently to address the demanding computational issues of Neural Architecture Search (NAS) in designing deep neural networks (DNNs), where each candidate architecture network only requires one iteration of backpropagation. The values obtained from proxies are used as predictions of architecture performance for downstream tasks. However, two significant drawbacks hinder the wide adoption of these efficient proxies: 1. they are not adaptive to various NAS search spaces and 2. they are not extensible to multi-modality downstream tasks. To address these two issues, we first propose an Extensible proxy (Eproxy) that utilizes self-supervised, few-shot training to achieve near-zero costs. A key component to our Eproxy's efficiency is the introduction of a barrier layer with randomly initialized frozen convolution parameters, which adds non-linearities to the optimization spaces so that Eproxy can discriminate the performance of architectures at an early stage. We further propose a Discrete Proxy Search (DPS) method to find the optimized training settings for Eproxy with only a handful of benchmarked architectures on the target tasks. Our extensive experiments confirm the effectiveness of both Eproxy and DPS. On the NDS-ImageNet search spaces, Eproxy+DPS achieves a higher average ranking correlation (Spearman r = 0.73) than the previous efficient proxy (Spearman r = 0.56). On the NAS-Bench-Trans-Micro search spaces with seven tasks, Eproxy+DPS delivers comparable performance with the early stopping method (146x faster). For the end-to-end task such as DARTS-ImageNet-1k, our method delivers better results than NAS performed on CIFAR-10 while only requiring one GPU hour with a single batch of CIFAR-10 images.
****
Zenseact Open Dataset: A Large-Scale and Diverse Multimodal Dataset for Autonomous Driving
Mina Alibeigi, William Ljungbergh, Adam Tonderski, Georg Hess, Adam Lilja, Carl Lindström, Daria Motorniuk, Junsheng Fu, Jenny Widahl, Christoffer Petersson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20178-20188
Existing datasets for autonomous driving (AD) often lack diversity and long-range capabilities, focusing instead on 360* perception and temporal reasoning. To address this gap, we introduce ZOD, a large-scale and diverse multimodal dataset collected over two years in various European countries, covering an area 9x that of existing datasets. ZOD boasts the highest range and resolution sensors among comparable datasets, coupled with detailed keyframe annotations for 2D and 3D objects (up to 245m), road instance/semantic segmentation, traffic sign recognition, and road classification. We believe that this unique combination will facilitate breakthroughs in long-range perception and multi-task learning. The dataset is composed of Frames, Sequences, and Drives, designed to encompass both data diversity and support for spatio-temporal learning, sensor fusion, localization, and mapping. Frames consist of 100k curated camera images with two seconds of other supporting sensor data, while the 1473 Sequences and 29 Drives include the entire sensor suite for 20 seconds and a few minutes, respectively. ZOD is the only AD dataset released under the permissive CC BY-SA 4.0 license, allowing for both research and commercial use. More information, and an extensive devkit, can be found at zod.zenseact.com.
****

MAAL: Multimodality-Aware Autoencoder-Based Affordance Learning for 3D Articulat
ed Objects
Yuanzhi Liang, Xiaohan Wang, Linchao Zhu, Yi Yang; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 217-227
Inferring affordance for 3D articulated objects is a challenging and practical p
roblem. It is a primary problem for applying robots to real-world scenarios. The
 exploration can be summarized as figuring out where to act and how to act. Corr
espondingly, the task mainly requires producing actionability scores, action pro
posals, and success likelihood scores according to the given 3D object informati
on and robotic information. Current works usually directly process multi-modal i
nputs with early fusion and apply critic networks to produce scores, which leads
 to insufficient multi-modal learning ability and inefficiently iterative traini
ng in multiple stages. This paper proposes a novel Multimodality-Aware Autoencod
er-based affordance Learning (MAAL) for the 3D object affordance problem. It is
an efficient pipeline, trained in one go, and only requires a few positive sampl
es in training data. More importantly, MAAL contains a MultiModal Energized Enco
der (MME) for better multi-modal learning. It comprehensively models all multi-m
odal inputs from 3D objects and robotic actions. Jointly considering information
 from multiple modalities, the encoder further learns interactions between robot
s and objects. MME empowers the better multi-modal learning ability for understa
nding object affordance. Experimental results and visualizations, based on a lar
ge-scale dataset PartNet-Mobility, show the effectiveness of MAAL in learning mu
lti-modal data and solving the 3D articulated object affordance problem.
****
Generalizable Decision Boundaries: Dualistic Meta-Learning for Open Set Domain G
eneralization
Xiran Wang, Jian Zhang, Lei Qi, Yinghuan Shi; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2023, pp. 11564-11573
Domain generalization (DG) is proposed to deal with the issue of domain shift, w
hich occurs when statistical differences exist between source and target domains
. However, most current methods do not account for a common realistic scenario w
here the source and target domains have different classes. To overcome this defi
ciency, open set domain generalization (OSDG) then emerges as a more practical s
etting to recognize unseen classes in unseen domains. An intuitive approach is t
o use multiple one-vs-all classifiers to define decision boundaries for each cla
ss and reject the outliers as unknown. However, the significant class imbalance
between positive and negative samples often causes the boundaries biased towards
 positive ones, resulting in misclassification for known samples in the unseen t
arget domain. In this paper, we propose a novel meta-learning-based framework ca
lled dualistic MEta-learning with joint DomaIn-Class matching (MEDIC), which con
siders gradient matching towards inter-domain and inter-class splits simultaneou
sly to find a generalizable boundary balanced for all tasks. Experimental result
s demonstrate that MEDIC not only outperforms previous methods in open set scena
rios, but also maintains competitive close set generalization ability at the sam
e time. Our code is available at https://github.com/zzwdx/MEDIC.
****
Benchmarking and Analyzing Robust Point Cloud Recognition: Bag of Tricks for Def
ending Adversarial Examples
Qiufan Ji, Lin Wang, Cong Shi, Shengshan Hu, Yingying Chen, Lichao Sun; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 4295-4304
Deep Neural Networks (DNNs) for 3D point cloud recognition are vulnerable to adv
ersarial examples, threatening their practical deployment. Despite the many rese
arch endeavors have been made to tackle this issue in recent years, the diversit
y of adversarial examples on 3D point clouds makes them more challenging to defe
nd against than those on 2D images. For examples, attackers can generate adversa
rial examples by adding, shifting, or removing points. Consequently, existing de
fense strategies are hard to counter unseen point cloud adversarial examples. In
 this paper, we first establish a comprehensive, and rigorous point cloud advers
arial robustness benchmark to evaluate adversarial robustness, which can provide

a detailed understanding of the effects of the defense and attack methods. We then collect existing defense tricks in point cloud adversarial defenses and then perform extensive and systematic experiments to identify an effective combination of these tricks. Furthermore, we propose a hybrid training augmentation methods that consider various types of point cloud adversarial examples to adversarial training, significantly improving the adversarial robustness. By combining these tricks, we construct a more robust defense framework achieving an average accuracy of 83.45% against various attacks, demonstrating its capability to enabling robust learners. Our codebase are open-sourced on https://github.com/qiufan319/benchmark_pc_attack.git
****
Weakly Supervised Referring Image Segmentation with Intra-Chunk and Inter-Chunk Consistency
Jungbeom Lee, Sungjin Lee, Jinseok Nam, Seunghak Yu, Jaeyoung Do, Tara Taghavi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21870-21881
Referring image segmentation (RIS) aims to localize the object in an image referred by a natural language expression. Most previous studies learn RIS with a large-scale dataset containing segmentation labels, but they are costly. We present a weakly supervised learning method for RIS that only uses readily available image-text pairs. We first train a visual-linguistic model for image-text matching and extract a visual saliency map through Grad-CAM to identify the image regions corresponding to each word. However, we found two major problems with Grad-CAM. First, it lacks consideration of critical semantic relationships between words. We tackle this problem by modeling the relationship between words through intra-chunk and inter-chunk consistency. Second, Grad-CAM identifies only small regions of the referred object, leading to low recall. Therefore, we refine the localization maps with self-attention in Transformer and unsupervised object shape prior. On three popular benchmarks (RefCOCO, RefCOCO+, G-Ref), our method significantly outperforms recent comparable techniques. We also show that our method is applicable to various levels of supervision and obtains better performance than recent methods.
****
Poincare ResNet
Max van Spengler, Erwin Berkhout, Pascal Mettes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5419-5428
This paper introduces an end-to-end residual network that operates entirely on the Poincare ball model of hyperbolic space. Hyperbolic learning has recently shown great potential for visual understanding, but is currently only performed in the penultimate layer(s) of deep networks. All visual representations are still learned through standard Euclidean networks. In this paper we investigate how to learn hyperbolic representations of visual data directly from the pixel-level. We propose Poincare ResNet, a hyperbolic counterpart of the celebrated residual network, starting from Poincare 2D convolutions up to Poincare residual connections. We identify three roadblocks for training convolutional networks entirely in hyperbolic space and propose a solution for each: (i) Current hyperbolic network initializations collapse to the origin, limiting their applicability in deeper networks. We provide an identity-based initialization that preserves norms over many layers. (ii) Residual networks rely heavily on batch normalization, which comes with expensive Frechet mean calculations in hyperbolic space. We introduce Poincare midpoint batch normalization as a faster and equally effective alternative. (iii) Due to the many intermediate operations in Poincare layers, the computation graphs of deep learning libraries blow up, limiting our ability to train on deep hyperbolic networks. We provide manual backward derivations of core hyperbolic operations to maintain manageable computation graphs.
****
Parameterized Cost Volume for Stereo Matching
Jiaxi Zeng, Chengtang Yao, Lidong Yu, Yuwei Wu, Yunde Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18347-18357
Stereo matching becomes computationally challenging when dealing with a large di

sparity range. Prior methods mainly alleviate the computation through dynamic cost volume by focusing on a local disparity space, but it requires many iterations to get close to the ground truth due to the lack of a global view. We find that the dynamic cost volume approximately encodes the disparity space as a single Gaussian distribution with a fixed and small variance at each iteration, which results in an inadequate global view over disparity space and a small update step at every iteration. In this paper, we propose a parameterized cost volume to encode the entire disparity space using multi-Gaussian distribution. The disparity distribution of each pixel is parameterized by weights, means, and variances. The means and variances are used to sample disparity candidates for cost computation, while the weights and means are used to calculate the disparity output. The above parameters are computed through a JS-divergence-based optimization, which is realized as a gradient descent update in a feed-forward differential module. Experiments show that our method speeds up the runtime of RAFT-Stereo by 4 15 times, achieving real-time performance and comparable accuracy.
****
SAFE: Sensitivity-Aware Features for Out-of-Distribution Object Detection
Samuel Wilson, Tobias Fischer, Feras Dayoub, Dimity Miller, Niko Sünderhauf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23565-23576
We address the problem of out-of-distribution (OOD) detection for the task of object detection. We show that residual convolutional layers with batch normalisation produce Sensitivity-Aware FEatures (SAFE) that are consistently powerful for distinguishing in-distribution from out-of-distribution detections. We extract SAFE vectors for every detected object, and train a multilayer perceptron on the surrogate task of distinguishing adversarially perturbed from clean in-distribution examples. This circumvents the need for realistic OOD training data, computationally expensive generative models, or retraining of the base object detector. SAFE outperforms the state-of-the-art OOD object detectors on multiple benchmarks by large margins, e.g. reducing the FPR95 by an absolute 30.6% from 48.3% to 17.7% on the OpenImages dataset.
****
SimFIR: A Simple Framework for Fisheye Image Rectification with Self-supervised Representation Learning
Hao Feng, Wendi Wang, Jiajun Deng, Wengang Zhou, Li Li, Houqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12418-12427
In fisheye images, rich distinct distortion patterns are regularly distributed in the image plane. These distortion patterns are independent of the visual content and provide informative cues for rectification. To make the best of such rectification cues, we introduce SimFIR, a simple framework for fisheye image rectification based on self-supervised representation learning. Technically, we first split a fisheye image into multiple patches and extract their representations with a Vision Transformer (ViT). To learn fine-grained distortion representations, we then associate different image patches with their specific distortion patterns based on the fisheye model, and further subtly design an innovative unified distortion-aware pretext task for their learning. The transfer performance on the downstream rectification task is remarkably boosted, which verifies the effectiveness of the learned representations. Extensive experiments are conducted, and the quantitative and qualitative results demonstrate the superiority of our method over the state-of-the-art algorithms as well as its strong generalization ability on real-world fisheye images.
****
Subclass-balancing Contrastive Learning for Long-tailed Recognition
Chengkai Hou, Jieyu Zhang, Haonan Wang, Tianyi Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5395-5407
Long-tailed recognition with imbalanced class distribution naturally emerges in practical machine learning applications. Existing methods such as data reweighing, resampling, and supervised contrastive learning enforce the class balance with a price of introducing imbalance between instances of head class and tail clas

s, which may ignore the underlying rich semantic substructures of the former and exaggerate the biases in the latter. We overcome these drawbacks by a novel "subclass-balancing contrastive learning (SBCL)" approach that clusters each head class into multiple subclasses of similar sizes as the tail classes and enforce representations to capture the two-layer class hierarchy between the original classes and their subclasses. Since the clustering is conducted in the representation space and updated during the course of training, the subclass labels preserve the semantic substructures of head classes. Meanwhile, it does not overemphasize tail class samples, so each individual instance contribute to the representation learning equally. Hence, our method achieves both the instance- and subclass-balance, while the original class labels are also learned through contrastive learning among subclasses from different classes. We evaluate SBCL over a list of long-tailed benchmark datasets and it achieves the state-of-the-art performance. In addition, we present extensive analyses and ablation studies of SBCL to verify its advantages.
****
Generalized Lightness Adaptation with Channel Selective Normalization
Mingde Yao, Jie Huang, Xin Jin, Ruikang Xu, Shenglong Zhou, Man Zhou, Zhiwei Xiong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10668-10679
Lightness adaptation is vital to the success of image processing to avoid unexpected visual deterioration, which covers multiple aspects, e.g., low-light image enhancement, image retouching, and inverse tone mapping. Existing methods typically work well on their trained lightness conditions but perform poorly in unknown ones due to their limited generalization ability. To address this limitation, we propose a novel generalized lightness adaptation algorithm that extends conventional normalization techniques through a channel filtering design, dubbed Channel Selective Normalization (CSNorm). The proposed CSNorm purposely normalizes the statistics of lightness-relevant channels and keeps other channels unchanged, so as to improve feature generalization and discrimination. To optimize CSNorm, we propose an alternating training strategy that effectively identifies lightness-relevant channels. The model equipped with our CSNorm only needs to be trained on one lightness condition and can be well generalized to unknown lightness conditions. Experimental results on multiple benchmark datasets demonstrate the effectiveness of CSNorm in enhancing the generalization ability for the existing lightness adaptation methods. Code is available at https://github.com/mdyao/CSNorm.
****
Omnidirectional Information Gathering for Knowledge Transfer-Based Audio-Visual Navigation
Jinyu Chen, Wenguan Wang, Si Liu, Hongsheng Li, Yi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10993-11003
Audio-visual navigation is an audio-targeted wayfinding task where a robot agent is entailed to travel a never-before-seen 3D environment towards the sounding source. In this article, we present ORAN, an omnidirectional audio-visual navigator based on cross-task navigation skill transfer. In particular, ORAN sharpens its two basic abilities for such challenging tasks, namely wayfinding and audio-visual information gathering. First, ORAN is trained with a confidence-aware cross-task policy distillation (CCPD) strategy. CCPD transfers the fundamental, point-to-point wayfinding skill that is well-trained on the large-scale PointGoal task to ORAN, to help ORAN better master audio-visual navigation with far fewer training samples. To improve the efficiency of knowledge transfer and address the domain gap, CCPD is made to be adaptive to the decision confidence of the teacher policy. Second, ORAN is equipped with an omnidirectional information gathering (OIG) mechanism, i.e., gleaning visual-acoustic observations from different directions before decision-making. As a result, ORAN yields more robust navigation behaviour. Taking CCPD and OIG together, ORAN significantly outperforms previous competitors. After the model ensemble, we got 1st in Soundspaces Challenge 2022, improving SPL and SR by 53% and 35% relatively. Our code will be released.
****

Multi-Scale Bidirectional Recurrent Network with Hybrid Correlation for Point Cl oud Based Scene Flow Estimation
Wencan Cheng, Jong Hwan Ko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10041-10050
Scene flow estimation provides the fundamental motion perception of a dynamic sc ene, which is of practical importance in many computer vision applications. In t his paper, we propose a novel multi-scale bidirectional recurrent architecture t hat iteratively optimizes the coarse-to-fine scene flow estimation. In each reso lution scale of estimation, a novel bidirectional gated recurrent unit is propos ed to bidirectionally and iteratively augment point features and produce progres sively optimized scene flow. The optimization of each iteration is integrated wi th the hybrid correlation that captures not only local correlation but also sema ntic correlation for more accurate estimation. Experimental results indicate tha t our proposed architecture significantly outperforms the existing state-of-the-art approaches on both FlyingThings3D and KITTI benchmarks while maintaining sup erior time efficiency. Codes and pre-trained models are publicly available at ht tps://github.com/cwc1260/MSBRN.
****
Dynamic Mesh-Aware Radiance Fields
Yi-Ling Qiao, Alexander Gao, Yiran Xu, Yue Feng, Jia-Bin Huang, Ming C. Lin; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202 3, pp. 385-396
Embedding polygonal mesh assets within photorealistic Neural Radience Fields (Ne RF) volumes, such that they can be rendered and their dynamics simulated in a ph ysically consistent manner with the NeRF, is under-explored from the system pers pective of integrating NeRF into the traditional graphics pipeline. This paper d esigns a two-way coupling between mesh and NeRF during rendering and simulation. We first review the light transport equations for both mesh and NeRF, then dist ill them into an efficient algorithm for updating radiance and throughput along a cast ray with an arbitrary number of bounces. To resolve the discrepancy betwe en the linear color space that the path tracer assumes and the sRGB color space that standard NeRF uses, we train NeRF with High Dynamic Range (HDR) images. We also present a strategy to estimate light sources and cast shadows on the NeRF. Finally, we consider how the hybrid surface-volumetric formulation can be effici ently integrated with a high-performance physics simulator that supports cloth, rigid and soft bodies. The full rendering and simulation system can be run on a GPU at interactive rates. We show that a hybrid system approach outperforms alte rnatives in visual realism for mesh insertion, because it allows realistic light transport from volumetric NeRF media onto surfaces, which affects the appearanc e of reflective/refractive surfaces and illumination of diffuse surfaces informe d by the dynamic scene.
****
Learning Support and Trivial Prototypes for Interpretable Image Classification
Chong Wang, Yuyuan Liu, Yuanhong Chen, Fengbei Liu, Yu Tian, Davis McCarthy, Hel en Frazer, Gustavo Carneiro; Proceedings of the IEEE/CVF International Conferenc e on Computer Vision (ICCV), 2023, pp. 2062-2072
Prototypical part network (ProtoPNet) methods have been designed to achieve inte rpretable classification by associating predictions with a set of training proto types, which we refer to as trivial prototypes because they are trained to lie f ar from the classification boundary in the feature space. Note that it is possib le to make an analogy between ProtoPNet and support vector machine (SVM) given t hat the classification from both methods relies on computing similarity with a s et of training points (i.e., trivial prototypes in ProtoPNet, and support vector s in SVM). However, while trivial prototypes are located far from the classifica tion boundary, support vectors are located close to this boundary, and we argue that this discrepancy with the well-established SVM theory can result in ProtoPN et models with inferior classification accuracy. In this paper, we aim to improv e the classification of ProtoPNet with a new method to learn support prototypes that lie near the classification boundary in the feature space, as suggested by the SVM theory. In addition, we target the improvement of classification results

with a new model, named ST-ProtoPNet, which exploits our support prototypes and the trivial prototypes to provide more effective classification. Experimental results on CUB-200-2011, Stanford Cars, and Stanford Dogs datasets demonstrate that ST-ProtoPNet achieves state-of-the-art classification accuracy and interpretability results. We also show that the proposed support prototypes tend to be better localised in the object of interest rather than in the background region. Code is available at https://github.com/cwangrun/ST-ProtoPNet.
****
Decoupled DETR: Spatially Disentangling Localization and Classification for Improved End-to-End Object Detection
Manyuan Zhang, Guanglu Song, Yu Liu, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6601-6610
The introduction of DETR represents a new paradigm for object detection.
However, its decoder conducts classification and box localization using shared queries and cross-attention layers, leading to suboptimal results. We observe that different regions of interest in the visual feature map are suitable for performing query classification and box localization tasks, even for the same object. Salient regions provide vital information for classification, while the boundaries around them are more favorable for box regression. Unfortunately, such spatial misalignment between these two tasks greatly hinders DETR's training.
Therefore, in this work, we focus on decoupling localization and classification tasks in DETR. To achieve this, we introduce a new design scheme called spatially decoupled DETR (SD-DETR), which includes a task-aware query generation module and a disentangled feature learning process.
We elaborately design the task-aware query initialization process and divide the cross-attention block in the decoder to allow the task-aware queries to match different visual regions.
Meanwhile, we also observe that the prediction misalignment problem for high classification confidence and precise localization exists, so we propose an alignment loss to further guide the spatially decoupled DETR training.
Through extensive experiments, we demonstrate that our approach achieves a significant improvement in MSCOCO datasets compared to previous work. For instance, we improve the performance of Conditional DETR by 4.5%. By spatially disentangling the two tasks, our method overcomes the misalignment problem and greatly improves the performance of DETR for object detection.
****
GIFD: A Generative Gradient Inversion Method with Feature Domain Optimization
Hao Fang, Bin Chen, Xuan Wang, Zhi Wang, Shu-Tao Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4967-4976
Federated Learning (FL) has recently emerged as a promising distributed machine learning framework to preserve clients' privacy, by allowing multiple clients to upload the gradients calculated from their local data to a central server. Recent studies find that the exchanged gradients also take the risk of privacy leakage, e.g., an attacker can invert the shared gradients and recover sensitive data against an FL system by leveraging pre-trained generative adversarial networks (GAN) as prior knowledge. However, performing gradient inversion attacks in the latent space of the GAN model limits their expression ability and generalizability. To tackle these challenges, we propose Gradient Inversion over Feature Domains (GIFD), which disassembles the GAN model and searches the feature domains of the intermediate layers. Instead of optimizing only over the initial latent code, we progressively change the optimized layer, from the initial latent space to intermediate layers closer to the output images. In addition, we design a regularizer to avoid unreal image generation by adding a small l1 ball constraint to the searching range. We also extend GIFD to the out-of-distribution (OOD) setting, which weakens the assumption that the training sets of GANs and FL tasks obey the same data distribution. Extensive experiments demonstrate that our method can achieve pixel-level reconstruction and is superior to the existing methods. Notably, GIFD also shows great generalizability under different defense strategy settings and batch sizes.
****

VLN-PETL: Parameter-Efficient Transfer Learning for Vision-and-Language Navigation

Yanyuan Qiao, Zheng Yu, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15443-15452

The performance of the Vision-and-Language Navigation (VLN) tasks has witnessed rapid progress recently thanks to the use of large pre-trained vision-and-language models. However, full fine-tuning the pre-trained model for every downstream VLN task is becoming costly due to the considerable model size. Recent research hotspot of Parameter-Efficient Transfer Learning (PETL) shows great potential in efficiently tuning large pre-trained models for the common CV and NLP tasks, which exploits the most of the representation knowledge implied in the pre-trained model while only tunes a minimal set of parameters. However, simply utilizing existing PETL methods for the more challenging VLN tasks may bring non-trivial degeneration to the performance. Therefore, we present the first study to explore PETL methods for VLN tasks and propose a VLN-specific PETL method named VLN-PETL. Specifically, we design two PETL modules: Historical Interaction Booster (HIB) and Cross-modal Interaction Booster (CIB). Then we combine these two modules with several existing PETL methods as the integrated VLN-PETL. Extensive experimental results on four mainstream VLN tasks (R2R, REVERIE, NDH, RxR) demonstrate the effectiveness of our proposed VLN-PETL, where VLN-PETL achieves comparable or even better performance to full fine-tuning and outperforms other PETL methods with promising margins. The source code is available at https://github.com/YanyuanQiao/VLN-PETL

****

Generalized Sum Pooling for Metric Learning

Yeti Z. Gürbüz, Ozan Sener, A. Aydin Alatan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5462-5473

A common architectural choice for deep metric learning is a convolutional neural network followed by global average pooling (GAP). Albeit simple, GAP is a highly effective way to aggregate information. One possible explanation for the effectiveness of GAP is considering each feature vector as representing a different semantic entity and GAP as a convex combination of them. Following this perspective, we generalize GAP and propose a learnable generalized sum pooling method (GSP). GSP improves GAP with two distinct abilities: i) the ability to choose a subset of semantic entities, effectively learning to ignore nuisance information, and ii) learning the weights corresponding to the importance of each entity. Formally, we propose an entropy-smoothed optimal transport problem and show that it is a strict generalization of GAP, i.e., a specific realization of the problem gives back GAP. We show that this optimization problem enjoys analytical gradients enabling us to use it as a direct learnable replacement for GAP. We further propose a zero-shot loss to ease the learning of GSP. We show the effectiveness of our method with extensive evaluations on 4 popular metric learning benchmarks. Code is available at: GSP-DML Framework

****

AlignDet: Aligning Pre-training and Fine-tuning in Object Detection

Ming Li, Jie Wu, Xionghui Wang, Chen Chen, Jie Qin, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6866-6876

The paradigm of large-scale pre-training followed by downstream fine-tuning has been widely employed in various object detection algorithms. In this paper, we reveal discrepancies in data, model, and task between the pre-training and fine-tuning procedure in existing practices, which implicitly limit the detector's performance, generalization ability, and convergence speed. To this end, we propose AlignDet, a unified pre-training framework that can be adapted to various existing detectors to alleviate the discrepancies. AlignDet decouples the pre-training process into two stages, i.e., image-domain and box-domain pre-training. The image-domain pre-training optimizes the detection backbone to capture holistic visual abstraction, and box-domain pre-training learns instance-level semantics and task-aware concepts to initialize the parts out of the backbone. By incorporating the self-supervised pre-trained backbones, we can pre-train all modules for

various detectors in an unsupervised paradigm. As depicted in Figure 1, extensive experiments demonstrate that AlignDet can achieve significant improvements across diverse protocols, such as detection algorithm, model backbone, data setting, and training schedule. For example, AlignDet improves FCOS by 5.3 mAP, RetinaNet by 2.1 mAP, Faster R-CNN by 3.3 mAP, and DETR by 2.3 mAP under fewer epochs.
****
Learning Continuous Exposure Value Representations for Single-Image HDR Reconstruction
Su-Kai Chen, Hung-Lin Yen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Wen-Hsiao Peng, Yen-Yu Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12990-13000
Deep learning is commonly used to produce impressive results in reconstructing HDR images from LDR images. LDR stack-based methods are used for single-image HDR reconstruction, generating an HDR image from a deep learning generated LDR stack. However, current methods generate the LDR stack with predetermined exposure values (EVs), which may limit the quality of HDR reconstruction. To address this, we propose the continuous exposure value representation (CEVR) model, which uses an implicit function to generate LDR images with arbitrary EVs, including those unseen during training. Our flexible approach generates a continuous stack with more images containing diverse EVs, significantly improving HDR reconstruction. We use a cycle training strategy to supervise the model in generating continuous EV LDR images without corresponding ground truths. Our CEVR model outperforms existing methods, as demonstrated by experimental results.
****
DREAM: Efficient Dataset Distillation by Representative Matching
Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, Yang You; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17314-17324
Dataset distillation aims to synthesize small datasets with little information loss from original large-scale ones for reducing storage and training costs. Recent state-of-the-art methods mainly constrain the sample synthesis process by matching synthetic images and the original ones regarding gradients, embedding distributions, or training trajectories. Although there are various matching objectives, currently the strategy for selecting original images is limited to naive random sampling. We argue that random sampling overlooks the evenness of the selected sample distribution, which may result in noisy or biased matching targets. Besides, the sample diversity is also not constrained by random sampling. These factors together lead to optimization instability in the distilling process and degrade the training efficiency. Accordingly, we propose a novel matching strategy named as Dataset distillation by REpresentAive Matching (DREAM), where only representative original images are selected for matching. DREAM is able to be easily plugged into popular dataset distillation frameworks and reduce the distilling iterations by more than 8 times without performance drop. Given sufficient training time, DREAM further provides significant improvements and achieves state-of-the-art performances.
****
MixSynthFormer: A Transformer Encoder-like Structure with Mixed Synthetic Self-attention for Efficient Human Pose Estimation
Yuran Sun, Alan William Dougherty, Zhuoying Zhang, Yi King Choi, Chuan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14884-14893
Human pose estimation in videos has wide-ranging practical applications across various fields, many of which require fast inference on resource-scarce devices, necessitating the development of efficient and accurate algorithms. Previous works have demonstrated the feasibility of exploiting motion continuity to conduct pose estimation using sparsely sampled frames with transformer-based models. However, these methods only consider the temporal relation while neglecting spatial attention, and the complexity of dot product self-attention calculations in transformers are quadratically proportional to the embedding size. To address these limitations, we propose MixSynthFormer, a transformer encoder-like model with M

LP-based mixed synthetic attention. By mixing synthesized spatial and temporal a
ttentions, our model incorporates inter-joint and inter-frame importance and can
 accurately estimate human poses in an entire video sequence from sparsely sampl
ed frames. Additionally, the flexible design of our model makes it versatile for
 other motion synthesis tasks. Our extensive experiments on 2D/3D pose estimatio
n, body mesh recovery, and motion prediction validate the effectiveness and effi
ciency of MixSynthFormer.
****
Focus on Your Target: A Dual Teacher-Student Framework for Domain-Adaptive Seman
tic Segmentation
Xinyue Huo, Lingxi Xie, Wengang Zhou, Houqiang Li, Qi Tian; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19027-1903
8
We study unsupervised domain adaptation (UDA) for semantic segmentation. Current
ly, a popular UDA framework lies in self-training which endows the model with tw
o-fold abilities: (i) learning reliable semantics from the labeled images in the
 source domain, and (ii) adapting to the target domain via generating pseudo lab
els on the unlabeled images. We find that, by decreasing/increasing the proporti
on of training samples from the target domain, the 'learning ability' is strengt
hened/weakened while the 'adapting ability' goes in the opposite direction, impl
ying a conflict between these two abilities, especially for a single model. To a
lleviate the issue, we propose a novel dual teacher-student (DTS) framework and
equip it with a bidirectional learning strategy. By increasing the proportion of
 target-domain data, the second teacher-student model learns to 'Focus on Your T
arget' while the first model is not affected. DTS is easily plugged into existin
g self-training approaches. In a standard UDA scenario (training on synthetic, l
abeled data and real, unlabeled data), DTS shows consistent gains over the basel
ines and sets new state-of-the-art results of 76.5% and 75.1% mIoUs on GTAv-City
scapes and SYNTHIA-Cityscapes, respectively. The implementation is available at
https://github.com/xinyuehuo/DTS.
****
Enhanced Meta Label Correction for Coping with Label Corruption
Mitchell Keren Taraday, Chaim Baskin; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2023, pp. 16295-16304
Deep Neural Networks (DNNs) have revolutionized visual classification tasks over
 the last decade.
 The training phase of deep-learning-based algorithms, however, often requires a
 vast amount of reliable annotated data.
 While reliability collecting such amount of labeled data usually yields to an e
xhaustive, expensive process,
 for many applications, acquiring massive datasets with imperfect annotations is
 straightforward.
 For instance, crawling search engines and online websites can generate a boatlo
ad amount of noisy labeled data. Hence, solving the problem of learning with noi
sy labels (LNL) is of paramount importance.
 Traditional LNL methods have successfully handled datasets with artificially in
jected noise, but they still fall short of adequately handling real-world noise.
 With the increasing use of meta-learning in the diverse fields of machine learn
ing, researchers have tried to leverage auxiliary small clean datasets to meta-c
orrect the training labels. Nonetheless, existing meta-label correction approach
es are not fully exploiting their potential. In this study, we propose EMLC, an
enhanced meta-label correction approach for the LNL problem.
 We re-examine the meta-learning process and introduce faster and more accurate
meta-gradient derivations. We propose a novel teacher architecture tailored expl
icitly for the LNL problem, equipped with novel training objectives.
 EMLC outperforms prior approaches and achieves state-of-the-art results in all
standard benchmarks.
 Notably, EMLC enhances the previous art on the noisy real-world dataset Clothin
g1M by 0.87%.
 Our publicly available code can be found at the following link: https://github.

com/iccv23anonymous/Enhanced-Meta-Label-Correction
****
Dense Text-to-Image Generation with Attention Modulation
Yunji Kim, Jiyoung Lee, Jin-Hwa Kim, Jung-Woo Ha, Jun-Yan Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7701-7711
Existing text-to-image diffusion models struggle to synthesize realistic images given dense captions, where each text prompt provides a detailed description for a specific image region. To address this, we propose DenseDiffusion, a training-free method that adapts a pre-trained text-to-image model to handle such dense captions while offering control over the scene layout. We first analyze the relationship between generated images' layouts and the pre-trained model's intermediate attention maps. Next, we develop an attention modulation method that guides objects to appear in specific regions according to layout guidance. Without requiring additional fine-tuning or datasets, we improve image generation performance given dense captions regarding both automatic and human evaluation scores. In addition, we achieve similar-quality visual results with models specifically trained with layout conditions.
****
HumanMAC: Masked Motion Completion for Human Motion Prediction
Ling-Hao Chen, JiaWei Zhang, Yewen Li, Yiren Pang, Xiaobo Xia, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9544-9555
Human motion prediction is a classical problem in computer vision and computer graphics, which has a wide range of practical applications. Previous effects achieve great empirical performance based on an encoding-decoding style. The methods of this style work by first encoding previous motions to latent representations and then decoding the latent representations into predicted motions. However, in practice, they are still unsatisfactory due to several issues, including complicated loss constraints, cumbersome training processes, and scarce switch of different categories of motions in prediction. In this paper, to address the above issues, we jump out of the foregoing style and propose a novel framework from a new perspective. Specifically, our framework works in a masked completion fashion. In the training stage, we learn a motion diffusion model that generates motions from random noise. In the inference stage, with a denoising procedure, we make motion prediction conditioning on observed motions to output more continuous and controllable predictions. The proposed framework enjoys promising algorithmic properties, which only needs one loss in optimization and is trained in an end-to-end manner. Additionally, it accomplishes the switch of different categories of motions effectively, which is significant in realistic tasks, e.g., the animation task. Comprehensive experiments on benchmarks confirm the superiority of the proposed framework. The project page is available at https://lhchen.top/Human-MAC.
****
Will Large-scale Generative Models Corrupt Future Datasets?
Ryuichiro Hataya, Han Bao, Hiromi Arai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20555-20565
Recently proposed large-scale text-to-image generative models such as DALLE 2, Midjourney, and StableDiffusion can generate high-quality and realistic images from users' prompts. Not limited to the research community, ordinary Internet users enjoy these generative models, and consequently, a tremendous amount of generated images have been shared on the Internet. Meanwhile, today's success of deep learning in the computer vision field owes a lot to images collected from the Internet. These trends lead us to a research question: "will such generated images impact the quality of future datasets and the performance of computer vision models positively or negatively?" This paper empirically answers this question by simulating contamination. Namely, we generate ImageNet-scale and COCO-scale datasets using a state-of-the-art generative model and evaluate models trained with "contaminated" datasets on various tasks, including image classification and image generation. Throughout experiments, we conclude that generated images negativ

ely affect downstream performance, while the significance depends on tasks and the amount of generated images. The generated datasets and the codes for experiments will be publicly released for future research. Generated datasets and source codes are available from https://github.com/moskomule/dataset-contamination.
****

SHACIRA: Scalable HAsh-grid Compression for Implicit Neural Representations
Sharath Girish, Abhinav Shrivastava, Kamal Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17513-17524
Implicit Neural Representations (INR) or neural fields have emerged as a popular framework to encode multimedia signals such as images and radiance fields while retaining high-quality. Recently, learnable feature grids such as Instant-NGP have allowed significant speed-up in the training as well as the sampling of INRs by replacing a large neural network with a multi-resolution look-up table of feature vectors and a much smaller neural network. However, these feature grids come at the expense of large memory consumption which can be a bottleneck for storage and streaming applications. In this work, we propose SHACIRA, a simple yet effective task-agnostic framework for compressing such feature grids with no additional post-hoc pruning/quantization stages. We reparameterize feature grids with quantized latent weights and apply entropy regularization in the latent space to achieve high levels of compression across various domains. Quantitative and qualitative results on diverse datasets consisting of images, videos, and radiance fields, show that our approach outperforms existing INR approaches without the need for any large datasets or domain-specific heuristics. Our project page is available at https://shacira.github.io
****

Prompt Switch: Efficient CLIP Adaptation for Text-Video Retrieval
Chaorui Deng, Qi Chen, Pengda Qin, Da Chen, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15648-15658
In text-video retrieval, recent works have benefited from the powerful learning capabilities of pre-trained text-image foundation models (e.g., CLIP) by adapting them to the video domain. A critical problem for them is how to effectively capture the rich semantics inside the video using the image encoder of CLIP. To tackle this, state-of-the-art methods adopt complex cross-modal modeling techniques to fuse the text information into video frame representations, which, however, incurs severe efficiency issues in large-scale retrieval systems as the video representations must be recomputed online for every text query. In this paper, we discard this problematic cross-modal fusion process and aim to learn semantically-enhanced representations purely from the video, so that the video representations can be computed offline and reused for different texts. Concretely, we first introduce a spatial-temporal "Prompt Cube" into the CLIP image encoder and iteratively switch it within the encoder layers to efficiently incorporate the global video semantics into frame representations. We then propose to apply an auxiliary video captioning objective to train the frame representations, which facilitates the learning of detailed video semantics by providing fine-grained guidance in the semantic space. With a naive temporal fusion strategy (i.e., mean-pooling) on the enhanced frame representations, we obtain state-of-the-art performances on three benchmark datasets, i.e., MSR-VTT, MSVD, and LSMDC.
****

Video Action Recognition with Attentive Semantic Units
Yifei Chen, Dapeng Chen, Ruijin Liu, Hao Li, Wei Peng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10170-10180
Visual-Language Models (VLMs) have significantly advanced video action recognition. Supervised by the semantics of action labels, recent works adapt the visual branch of VLMs to learn video representations. Despite the effectiveness proved by these works, we believe that the potential of VLMs has yet to be fully harnessed. In light of this, we exploit the semantic units (SU) hiding behind the action labels and leverage their correlations with fine-grained items in frames for more accurate action recognition. SUs are entities extracted from the language descriptions of the entire action set, including body parts, objects, scenes, and motions. To further enhance the alignments between visual contents and the SUs,

we introduce a multi-region module (MRA) to the visual branch of the VLM. The M
RA allows the perception of region-aware visual features beyond the original glo
bal feature. Our method adaptively attends to and selects relevant SUs with visu
al features of frames. With a cross-modal decoder, the selected SUs serve to dec
ode spatiotemporal video representations. In summary, the SUs as the medium can
boost discriminative ability and transferability. Specifically, in fully-supervi
sed learning, our method achieved 87.8% top-1 accuracy on Kinetics-400. In K=2 f
ew-shot experiments, our method surpassed the previous state-of-the-art by +7.1%
and +15.0% on HMDB-51 and UCF-101, respectively.
****
Sentence Attention Blocks for Answer Grounding
Seyedalireza Khoshsirat, Chandra Kambhamettu; Proceedings of the IEEE/CVF Intern
ational Conference on Computer Vision (ICCV), 2023, pp. 6080-6090
Answer grounding is the task of locating relevant visual evidence for the Visual
Question Answering task. While a wide variety of attention methods have been in
troduced for this task, they suffer from the following three problems: designs t
hat do not allow the usage of pre-trained networks and do not benefit from large
data pre-training, custom designs that are not based on well-grounded previous
designs, therefore limiting the learning power of the network, or complicated de
signs that make it challenging to re-implement or improve them. In this paper, w
e propose a novel architectural block, which we term Sentence Attention Block, t
o solve these problems. The proposed block re-calibrates channel-wise image feat
ure-maps by explicitly modeling inter-dependencies between the image feature-map
s and sentence embedding. We visually demonstrate how this block filters out irr
elevant feature-maps channels based on sentence embedding. We start our design w
ith a well-known attention method, and by making minor modifications, we improve
the results to achieve state-of-the-art accuracy. The flexibility of our method
makes it easy to use different pre-trained backbone networks, and its simplicit
y makes it easy to understand and be re-implemented. We demonstrate the effectiv
eness of our method on the TextVQA-X, VQS, VQA-X, and VizWiz-VQA-Grounding datas
ets. We perform multiple ablation studies to show the effectiveness of our desig
n choices.
****
Scanning Only Once: An End-to-end Framework for Fast Temporal Grounding in Long
Videos
Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, Deli Zha
o; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 13767-13777
Video temporal grounding aims to pinpoint a video segment that matches the query
description. Despite the recent advance in short-form videos (e.g., in minutes)
, temporal grounding in long videos (e.g., in hours) is still at its early stage
. To address this challenge, a common practice is to employ a sliding window, ye
t can be inefficient and inflexible due to the limited number of frames within t
he window. In this work, we propose an end-to-end framework for fast temporal gr
ounding, which is able to model an hours-long video with one-time network execut
ion. Our pipeline is formulated in a coarse-to-fine manner, where we first extra
ct context knowledge from non-overlapped video clips (i.e., anchors), and then s
upplement the anchors that highly response to the query with detailed content kn
owledge. Besides the remarkably high pipeline efficiency, another advantage of o
ur approach is the capability of capturing long-range temporal correlation, than
ks to modeling the entire video as a whole, and hence facilitates more accurate
grounding. Experimental results suggest that, on the long-form video datasets MA
D and Ego4d, our method significantly outperforms state-of-the-arts, and achieve
s 14.6x / 102.8x higher efficiency respectively.
****
A Low-Shot Object Counting Network With Iterative Prototype Adaptation
Nikola ■uki■, Alan Luke■i■, Vitjan Zavrtanik, Matej Kristan; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18872-188
81
We consider low-shot counting of arbitrary semantic categories in the image usin

g only few annotated exemplars (few-shot) or no exemplars (no-shot). The standar
d few-shot pipeline follows extraction of appearance queries from exemplars and
matching them with image features to infer the object counts. Existing methods e
xtract queries by feature pooling which neglects the shape information (e.g., si
ze and aspect) and leads to a reduced object localization accuracy and count est
imates. We propose a Low-shot Object Counting network with iterative prototype A
daptation (LOCA). Our main contribution is the new object prototype extraction m
odule, which iteratively fuses the exemplar shape and appearance information wit
h image features. The module is easily adapted to zero-shot scenarios, enabling
LOCA to cover the entire spectrum of low-shot counting problems. LOCA outperform
s all recent state-of-the-art methods on FSC147 benchmark by 20-30% in RMSE on o
ne-shot and few-shot and achieves state-of-the-art on zero-shot scenarios, while
 demonstrating better generalization capabilities. The code and models are avail
able here: https://github.com/djukicn/loca.
****
Towards Fairness-aware Adversarial Network Pruning
Lei Zhang, Zhibo Wang, Xiaowei Dong, Yunhe Feng, Xiaoyi Pang, Zhifei Zhang, Kui
Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 5168-5177
Network pruning aims to compress models while minimizing loss in accuracy. With
the increasing focus on bias in AI systems, the bias inheriting or even magnific
ation nature of traditional network pruning methods has raised a new perspective
 towards fairness-aware network pruning. Straightforward pruning plus debias met
hods and recent designs for monitoring disparities of demographic attributes dur
ing pruning have endeavored to enhance fairness in pruning. However, neither sim
ple assembling of two tasks nor specifically designed pruning strategies could a
chieve the optimal trade-off among pruning ratio, accuracy, and fairness. This p
aper proposes an end-to-end learnable framework for fairness-aware network pruni
ng, which optimizes both pruning and debias tasks jointly by adversarial trainin
g against those final evaluation metrics like accuracy for pruning, and disparat
e impact (DI) and equalized odds (DEO) for fairness. In other words, our fairnes
s-aware adversarial pruning method would learn to prune without any handcraft ru
les. Therefore, our approach could flexibly adapt to variate network structures.
 Exhaustive experimentation demonstrates the generalization capacity of our appr
oach, as well as superior performance on pruning and debias simultaneously. To h
ighlight, the proposed method could preserve the SOTA pruning performance while
significantly improving fairness by around 50% as compared to traditional prunin
g methods.
****
VoroMesh: Learning Watertight Surface Meshes with Voronoi Diagrams
Nissim Maruani, Roman Klokov, Maks Ovsjanikov, Pierre Alliez, Mathieu Desbrun; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 14565-14574
In stark contrast to the case of images, finding a concise, learnable discrete r
epresentation of 3D surfaces remains a challenge. In particular, while polygon m
eshes are arguably the most common surface representation used in geometry proce
ssing, their irregular and combinatorial structure often make them unsuitable fo
r learning-based applications. In this work, we present VoroMesh, a novel and di
fferentiable of watertight 3D shape surfaces. From a set of 3D points (called ge
nerators) and their associated occupancy, we define our boundary representation
through the Voronoi diagram of the generators as the subset of Voronoi faces who
se two associated (equidistant) generators are of opposite occupancy: the result
ing polygon mesh forms a watertight approximation of the target shape's boundary
. To learn the position of the generators, we propose a novel loss function, dub
bed VoroLoss, that minimizes the distance from ground truth surface samples to t
he closest faces of the Voronoi diagram which does not require an explicit const
ruction of the entire Voronoi diagram. A direct optimization of the Voroloss to
obtain generators on the Thingi32 dataset demonstrates the geometric efficiency
of our representation compared to axiomatic meshing algorithms and recent learni
ng-based mesh representations. We further use VoroMesh in a learning-based mesh

prediction task from input SDF grids on the ABC dataset, and show comparable per
formance to state-of-the-art methods while guaranteeing closed output surfaces f
ree of self-intersections.
****
Breaking Temporal Consistency: Generating Video Universal Adversarial Perturbati
ons Using Image Models
Hee-Seon Kim, Minji Son, Minbeom Kim, Myung-Joon Kwon, Changick Kim; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4
325-4334
As video analysis using deep learning models becomes more widespread, the vulner
ability of such models to adversarial attacks is becoming a pressing concern.
 In particular, Universal Adversarial Perturbation (UAP) poses a significant thr
eat, as a single perturbation can mislead deep learning models on entire dataset
s.
 We propose a novel video UAP using image data and image model.
 This enables us to take advantage of the rich image data and image model-based
studies available for video applications.
 However, there is a challenge that image models are limited in their ability to
 analyze the temporal aspects of videos, which is crucial for a successful video
 attack.
 To address this challenge, we introduce the Breaking Temporal Consistancy (BTC)
 method, which is the first attempt to incorporate temporal information into vid
eo attacks using image models.
 We aim to generate adversarial videos that have opposite patterns to the origin
al. Specifically, BTC-UAP minimizes the feature similarity between neighboring f
rames in videos.
 Our approach is simple but effective at attacking unseen video models.
 Additionally, it is applicable to videos of varying lengths and invariant to te
mporal shifts.
 Our approach surpasses existing methods in terms of effectiveness on various da
tasets, including ImageNet, UCF-101, and Kinetics-400.
****
Smoothness Similarity Regularization for Few-Shot GAN Adaptation
Vadim Sushko, Ruyu Wang, Juergen Gall; Proceedings of the IEEE/CVF International
 Conference on Computer Vision (ICCV), 2023, pp. 7073-7082
The task of few-shot GAN adaptation aims to adapt a pre-trained GAN model to a s
mall dataset with very few training images. While existing methods perform well
when the dataset for pre-training is structurally similar to the target dataset,
 the approaches suffer from training instabilities or memorization issues when t
he objects in the two domains have a very different structure. To mitigate this
limitation, we propose a new smoothness similarity regularization that transfers
 the inherently learned smoothness of the pre-trained GAN to the few-shot target
 domain even if the two domains are very different. We evaluate our approach by
adapting an unconditional and a class-conditional GAN to diverse few-shot target
 domains. Our proposed method significantly outperforms prior few-shot GAN adapt
ation methods in the challenging case of structurally dissimilar source-target d
omains, while performing on par with the state of the art for similar source-tar
get domains.
****
Distilling Coarse-to-Fine Semantic Matching Knowledge for Weakly Supervised 3D V
isual Grounding
Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong
 Yin, Zhou Zhao; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 2662-2671
3D visual grounding involves finding a target object in a 3D scene that correspo
nds to a given sentence query. Although many approaches have been proposed and a
chieved impressive performance, they all require dense object-sentence pair anno
tations in 3D point clouds, which are both time-consuming and expensive. To addr
ess the problem that fine-grained annotated data is difficult to obtain, we prop
ose to leverage weakly supervised annotations to learn the 3D visual grounding m

odel, i.e., only coarse scene-sentence correspondences are used to learn object-sentence links. To accomplish this, we design a novel semantic matching model that analyzes the semantic similarity between object proposals and sentences in a coarse-to-fine manner. Specifically, we first extract object proposals and coarsely select the top-K candidates based on feature and class similarity matrices. Next, we reconstruct the masked keywords of the sentence using each candidate one by one, and the reconstructed accuracy finely reflects the semantic similarity of each candidate to the query. Additionally, we distill the coarse-to-fine semantic matching knowledge into a typical two-stage 3D visual grounding model, which reduces inference costs and improves performance by taking full advantage of the well-studied structure of the existing architectures. We conduct extensive experiments on ScanRefer, Nr3D, and Sr3D, which demonstrate the effectiveness of our proposed method.
****
What does CLIP know about a red circle? Visual prompt engineering for VLMs
Aleksandar Shtedritski, Christian Rupprecht, Andrea Vedaldi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11987-11997
Large-scale Vision-Language Models, such as CLIP, learn powerful image-text representations that have found numerous applications, from zero-shot classification to text-to-image generation. Despite that, their capabilities for solving novel discriminative tasks via prompting fall behind those of large language models, such as GPT-3. Here we explore the idea of visual prompt engineering for solving computer vision tasks beyond classification by editing in image space instead of text. In particular, we discover an emergent ability of CLIP, where, by simply drawing a red circle around an object, we can direct the model's attention to that region, while also maintaining global information. We show the power of this simple approach by achieving state-of-the-art in zero-shot referring expressions comprehension and strong performance in keypoint localization tasks. Finally, we draw attention to some potential ethical concerns of large language-vision models.
****
MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation
Najmeh Sadoughi, Xinyu Li, Avijit Vajpayee, David Fan, Bing Shuai, Hector Santos-Villalobos, Vimal Bhat, Rohith MV; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23331-23340
Previous research has studied the task of segmenting cinematic videos into scenes and into narrative acts. However, these studies have overlooked the essential task of multimodal alignment and fusion for effectively and efficiently processing long-form videos (>60min). In this paper, we introduce Multimodal alignmEnt aGgregation and distillAtion (MEGA) for cinematic long-video segmentation. MEGA tackles the challenge by leveraging multiple media modalities. The method coarsely aligns inputs of variable lengths and different modalities with alignment positional encoding. To maintain temporal synchronization while reducing computation, we further introduce an enhanced bottleneck fusion layer which uses temporal alignment. Additionally, MEGA employs a novel contrastive loss to synchronize and transfer labels across modalities, enabling act segmentation from labeled synopsis sentences on video shots. Our experimental results show that MEGA outperforms state-of-the-art methods on MovieNet dataset for scene segmentation (with an Average Precision improvement of +1.19%) and on TRIPOD dataset for act segmentation (with a Total Agreement improvement of +5.51%).
****
DiffRate : Differentiable Compression Rate for Efficient Vision Transformers
Mengzhao Chen, Wenqi Shao, Peng Xu, Mingbao Lin, Kaipeng Zhang, Fei Chao, Rongrong Ji, Yu Qiao, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17164-17174
Token compression aims to speed up large-scale vision transformers (e.g. ViTs) by pruning (dropping) or merging tokens. It is an important but challenging task. Although recent advanced approaches achieved great success, they need to carefu

lly handcraft a compression rate (i.e. number of tokens to remove), which is ted ious and leads to sub-optimal performance. To tackle this problem, we propose Di fferentiable Compression Rate (DiffRate), a novel token compression method that has several appealing properties prior arts do not have. First, DiffRate enables propagating the loss function's gradient onto the compression ratio, which is c onsidered as a non-differentiable hyperparameter in previous work. In this case, different layers can automatically learn different compression rates layer-wise ly without extra overhead. Second, token pruning and merging can be naturally pe rformed simultaneously in DiffRate, while they were isolated in previous works. Third, extensive experiments demonstrate that DiffRate achieves state-of-the-art performance. For example, by applying the learned layer-wise compression rates to an off-the-shelf ViT-H (MAE) model, we achieve a 40% FLOPs reduction and a 1. 5x throughput improvement, with a minor accuracy drop of 0.16% on ImageNet witho ut fine-tuning, even outperforming previous methods with fine-tuning. Codes and models are available at https://github.com/OpenGVLab/DiffRate.
****
zPROBE: Zero Peek Robustness Checks for Federated Learning
Zahra Ghodsi, Mojan Javaheripi, Nojan Sheybani, Xinqiao Zhang, Ke Huang, Farinaz Koushanfar; Proceedings of the IEEE/CVF International Conference on Computer Vi sion (ICCV), 2023, pp. 4860-4870
Privacy-preserving federated learning allows multiple users to jointly train a m odel with coordination of a central server. The server only learns the final agg regation result, thereby preventing leakage of the users' (private) training dat a from the individual model updates. However, keeping the individual updates pri vate allows malicious users to degrade the model accuracy without being detected , also known as Byzantine attacks. Best existing defenses against Byzantine work ers rely on robust rank-based statistics, e.g., setting robust bounds via the me dian of updates, to find malicious updates. However, implementing privacy-preser ving rank-based statistics, especially median-based, is nontrivial and unscalabl e in the secure domain, as it requires sorting of all individual updates. We est ablish the first private robustness check that uses high break point rank-based statistics on aggregated model updates. By exploiting randomized clustering, we significantly improve the scalability of our defense without compromising privac y. We leverage the derived statistical bounds in zero-knowledge proofs to detect and remove malicious updates without revealing the private user updates. Our no vel framework, zPROBE, enables Byzantine resilient and secure federated learning . We show the effectiveness of zPROBE on several computer vision benchmarks. Emp irical evaluations demonstrate that zPROBE provides a low overhead solution to d efend against state-of-the-art Byzantine attacks while preserving privacy.
****
LoLep: Single-View View Synthesis with Locally-Learned Planes and Self-Attention Occlusion Inference
Cong Wang, Yu-Ping Wang, Dinesh Manocha; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2023, pp. 10841-10851
We propose a novel method, LoLep, which regresses Locally-Learned planes from a single RGB image to represent scenes accurately, thus generating better novel vi ews. Without the depth information, regressing appropriate plane locations is a challenging problem. To solve this issue, we pre-partition the disparity space i nto bins and design a disparity sampler to regress local offsets for multiple pl anes in each bin. However, only using such a sampler makes the network not conve rgent; we further propose two optimizing strategies that combine with different disparity distributions of datasets and propose an occlusion-aware reprojection loss as a simple yet effective geometric supervision technique. We also introduc e a self-attention mechanism to improve occlusion inference and present a Block-Sampling Self-Attention (BS-SA) module to address the problem of applying self-a ttention to large feature maps. We demonstrate the effectiveness of our approach and generate state-of-the-art results on different datasets. Compared to MINE, our approach has an LPIPS reduction of 4.8% 9.0% and an RV reduction of 74.9% 83 .5%. We also evaluate the performance on real-world images and demonstrate the b enefits. We will release the source code at the time of publication.

****
Multi-Modal Continual Test-Time Adaptation for 3D Semantic Segmentation
Haozhi Cao, Yuecong Xu, Jianfei Yang, Pengyu Yin, Shenghai Yuan, Lihua Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18809-18819
Continual Test-Time Adaptation (CTTA) generalizes conventional Test-Time Adaptation (TTA) by assuming that the target domain is dynamic over time rather than stationary. In this paper, we explore Multi-Modal Continual Test-Time Adaptation (MM-CTTA) as a new extension of CTTA for 3D semantic segmentation. The key to MM-CTTA is to adaptively attend to the reliable modality while avoiding catastrophic forgetting during continual domain shifts, which is out of the capability of previous TTA or CTTA methods. To fulfill this gap, we propose an MM-CTTA method called Continual Cross-Modal Adaptive Clustering (CoMAC) that addresses this task from two perspectives. On one hand, we propose an adaptive dual-stage mechanism to generate reliable cross-modal predictions by attending to the reliable modality based on the class-wise feature-centroid distance in the latent space. On the other hand, to perform test-time adaptation without catastrophic forgetting, we design class-wise momentum queues that capture confident target features for adaptation while stochastically restoring pseudo-source features to revisit source knowledge. We further introduce two new benchmarks to facilitate the exploration of MM-CTTA in the future. Our experimental results show that our method achieves state-of-the-art performance on both benchmarks. Visit our project website at https://sites.google.com/view/mmcotta.
****
Exploring Positional Characteristics of Dual-Pixel Data for Camera Autofocus
Myungsub Choi, Hana Lee, Hyong-euk Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13158-13168
In digital photography, autofocus is a key feature that aids high-quality image capture, and modern approaches use the phase patterns arising from dual-pixel sensors as important focus cues. However, dual-pixel data is prone to multiple error sources in its image capturing process, including lens shading or distortions due to the inherent optical characteristics of the lens. We observe that, while these degradations are hard to model using prior knowledge, they are correlated with the spatial position of the pixels within the image sensor area, and we propose a learning-based autofocus model with positional encodings (PE) to capture these patterns. Specifically, we introduce RoI-PE, which encodes the spatial position of our focusing region-of-interest (RoI) on the imaging plane. Learning with RoI-PE allows the model to be more robust to spatially-correlated degradations. In addition, we also propose to encode the current focal position of lens as lens-PE, which allows us to significantly reduce the computational complexity of the autofocus model. Experimental results clearly demonstrate the effectiveness of using the proposed position encodings for automatic focusing based on dual-pixel data.
****
Heterogeneous Forgetting Compensation for Class-Incremental Learning
Jiahua Dong, Wenqi Liang, Yang Cong, Gan Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11742-11751
Class-incremental learning (CIL) has achieved remarkable successes in learning new classes consecutively while overcoming catastrophic forgetting on old categories. However, most existing CIL methods unreasonably assume that all old categories have the same forgetting pace, and neglect negative influence of forgetting heterogeneity among different old classes on forgetting compensation. To surmount the above challenges, we develop a novel Heterogeneous Forgetting Compensation (HFC) model, which can resolve heterogeneous forgetting of easy-to-forget and hard-to-forget old categories from both representation and gradient aspects. Specifically, we design a task-semantic aggregation block to alleviate heterogeneous forgetting from representation aspect. It aggregates local category information within each task to learn task-shared global representations. Moreover, we develop two novel plug-and-play losses: a gradient-balanced forgetting compensation loss and a gradient-balanced relation distillation loss to alleviate forgetting

from gradient aspect. They consider gradient-balanced compensation to rectify forgetting heterogeneity of old categories and heterogeneous relation consistency. Experiments on several representative datasets illustrate effectiveness of our HFC model. The code is available at https://github.com/JiahuaDong/HFC.
****
FemtoDet: An Object Detection Baseline for Energy Versus Performance Tradeoffs
Peng Tu, Xu Xie, Guo Ai, Yuexiang Li, Yawen Huang, Yefeng Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13318-13327
Efficient detectors for edge devices are often optimized for parameters or speed count metrics, which remain in weak correlation with the energy of detectors.
However, some vision applications of convolutional neural networks, such as always-on surveillance cameras, are critical for energy constraints.
This paper aims to serve as a baseline by designing detectors to reach tradeoffs between energy and performance from two perspectives:
1) We extensively analyze various CNNs to identify low-energy architectures, including selecting activation functions, convolutions operators, and feature fusion structures on necks. These underappreciated details in past work seriously affect the energy consumption of detectors;
2) To break through the dilemmatic energy-performance problem, we propose a balanced detector driven by energy using discovered low-energy components named FemtoDet.
In addition to the novel construction, we improve FemtoDet by considering convolutions and training strategy optimizations.
Specifically, we develop a new instance boundary enhancement (IBE) module for convolution optimization to overcome the contradiction between the limited capacity of CNNs and detection tasks in diverse spatial representations, and propose a recursive warm-restart (RecWR) for optimizing training strategy to escape the sub-optimization of light-weight detectors by considering the data shift produced in popular augmentations.
As a result, FemtoDet with only 68.77k parameters achieves a competitive score of 46.3 AP50 on PASCAL VOC and 1.11 W & 64.47 FPS on Qualcomm Snapdragon 865 CPU platforms.
Extensive experiments on COCO and TJU-DHD datasets indicate that the proposed method achieves competitive results in diverse scenes.
****
Generative Prompt Model for Weakly Supervised Object Localization
Yuzhong Zhao, Qixiang Ye, Weijia Wu, Chunhua Shen, Fang Wan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6351-6361
Weakly supervised object localization (WSOL) remains challenging when learning object localization models from image category labels. Conventional methods that discriminatively train activation models ignore representative yet less discriminative object parts. In this study, we propose a generative prompt model (GenPromp), defining the first generative pipeline to localize less discriminative object parts by formulating WSOL as a conditional image denoising procedure. During training, GenPromp converts image category labels to learnable prompt embeddings which are fed to a generative model to conditionally recover the input image with noise and learn representative embeddings. During inference, GenPromp combines the representative embeddings with discriminative embeddings (queried from an off-the-shelf vision-language model) for both representative and discriminative capacity. The combined embeddings are finally used to generate multi-scale high-quality attention maps, which facilitate localizing full object extent. Experiments on CUB-200-2011 and ILSVRC show that GenPromp respectively outperforms the best discriminative models by 5.2% and 5.6% (Top-1 Loc), setting a solid baseline for WSOL with the generative model. Code is available at https://github.com/callsys/GenPromp.
****
ActFormer: A GAN-based Transformer towards General Action-Conditioned 3D Human Motion Generation
Liang Xu, Ziyang Song, Dongliang Wang, Jing Su, Zhicheng Fang, Chenjing Ding, We

ihao Gan, Yichao Yan, Xin Jin, Xiaokang Yang, Wenjun Zeng, Wei Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2228-2238

We present a GAN-based Transformer for general action-conditioned 3D human motion generation, including not only single-person actions but also multi-person interactive actions. Our approach consists of a powerful Action-conditioned motion TransFormer (ActFormer) under a GAN training scheme, equipped with a Gaussian Process latent prior. Such a design combines the strong spatio-temporal representation capacity of Transformer, superiority in generative modeling of GAN, and inherent temporal correlations from the latent prior. Furthermore, ActFormer can be naturally extended to multi-person motions by alternately modeling temporal correlations and human interactions with Transformer encoders. To further facilitate research on multi-person motion generation, we introduce a new synthetic dataset of complex multi-person combat behaviors. Extensive experiments on NTU-13, NTU RGB+D 120, BABEL and the proposed combat dataset show that our method can adapt to various human motion representations and achieve superior performance over the state-of-the-art methods on both single-person and multi-person motion generation tasks, demonstrating a promising step towards a general human motion generator.

****

Hiding Visual Information via Obfuscating Adversarial Perturbations

Zhigang Su, Dawei Zhou, Nannan Wang, Decheng Liu, Zhen Wang, Xinbo Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4356-4366

Growing leakage and misuse of visual information raise security and privacy concerns, which promotes the development of information protection. Existing adversarial perturbations-based methods mainly focus on the de-identification against deep learning models. However, the inherent visual information of the data has not been well protected. In this work, inspired by the Type-I adversarial attack, we propose an Adversarial Visual Information Hiding (AVIH) method to protect the visual privacy of data. Specifically, the method generates obfuscating adversarial perturbations to obscure the visual information of the data. Meanwhile, it maintains the hidden objectives to be correctly predicted by models. In addition, our method does not modify the parameters of the applied model, which makes it flexible for different scenarios. Experimental results on the recognition and classification tasks demonstrate that the proposed method can effectively hide visual information and hardly affect the performances of models. The code is available at https://github.com/suzhigangssz/AVIH.

****

Category-aware Allocation Transformer for Weakly Supervised Object Localization

Zhiwei Chen, Jinren Ding, Liujuan Cao, Yunhang Shen, Shengchuan Zhang, Guannan Jiang, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6643-6652

Weakly supervised object localization (WSOL) aims to localize objects based on only image-level labels as supervision. Recently, transformers have been introduced into WSOL, yielding impressive results. The self-attention mechanism and multilayer perceptron structure in transformers preserve long-range feature dependency, facilitating complete localization of the full object extent. However, current transformer-based methods predict bounding boxes using category-agnostic attention maps, which may lead to confused and noisy object localization. To address this issue, we propose a novel Category-aware Allocation TRansformer (CATR) that learns category-aware representations for specific objects and produces corresponding category-aware attention maps for object localization. First, we introduce a Category-aware Stimulation Module (CSM) to induce learnable category biases for self-attention maps, providing auxiliary supervision to guide the learning of more effective transformer representations. Second, we design an Object Constraint Module (OCM) to refine the object regions for the category-aware attention maps in a self-supervised manner. Extensive experiments on the CUB-200-2011 and ILSVRC datasets demonstrate that the proposed CATR achieves significant and consistent performance improvements over competing approaches.

****
Domain Specified Optimization for Deployment Authorization
Haotian Wang, Haoang Chi, Wenjing Yang, Zhipeng Lin, Mingyang Geng, Long Lan, Jing Zhang, Dacheng Tao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5095-5105
This paper explores Deployment Authorization (DPA) as a means of restricting the generalization capabilities of vision models on certain domains to protect intellectual property. Nevertheless, the current advancements in DPA are predominantly confined to fully supervised settings. Such settings require the accessibility of annotated images from any unauthorized domain, rendering the DPA approach impractical for real-world applications due to its exorbitant costs. To address this issue, we propose Source-Only Deployment Authorization (SDPA), which assumes that only authorized domains are accessible during training phases, and the model's performance on unauthorized domains must be suppressed in inference stages. Drawing inspiration from distributional robust statistics, we present a lightweight method called Domain-Specified Optimization (DSO) for SDPA that degrades the model's generalization over a divergence ball. DSO comes with theoretical guarantees on the convergence property and its authorization performance. As a complementary of SDPA, we also propose Target-Combined Deployment Authorization (TPDA), where unauthorized domains are partially accessible, and simplify the DSO method to a perturbation operation on the pseudo predictions, referred to as Target-Dependent Domain-Specified Optimization (TDSO). We demonstrate the effectiveness of our proposed DSO and TDSO methods through extensive experiments on six image benchmarks, achieving dominant performance on both SDPA and TDPA settings.
****
Iterative Prompt Learning for Unsupervised Backlit Image Enhancement
Zhexin Liang, Chongyi Li, Shangchen Zhou, Ruicheng Feng, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8094-8103
We propose a novel unsupervised backlit image enhancement method, abbreviated as CLIP-LIT, by exploring the potential of Contrastive Language-Image Pre-Training (CLIP) for pixel-level image enhancement. We show that the open-world CLIP prior not only aids in distinguishing between backlit and well-lit images, but also in perceiving heterogeneous regions with different luminance, facilitating the optimization of the enhancement network. Unlike high-level and image manipulation tasks, directly applying CLIP to enhancement tasks is non-trivial, owing to the difficulty in finding accurate prompts. To solve this issue, we devise a prompt learning framework that first learns an initial prompt pair by constraining the text-image similarity between the prompt (negative/positive sample) and the corresponding image (backlit image/well-lit image) in the CLIP latent space. Then, we train the enhancement network based on the text-image similarity between the enhanced result and the initial prompt pair. To further improve the accuracy of the initial prompt pair, we iteratively fine-tune the prompt learning framework to reduce the distribution gaps between the backlit images, enhanced results, and well-lit images via rank learning, boosting the enhancement performance. Our method alternates between updating the prompt learning framework and enhancement network until visually pleasing results are achieved. Extensive experiments demonstrate that our method outperforms state-of-the-art methods in terms of visual quality and generalization ability, without requiring any paired data.
****
UMIFormer: Mining the Correlations between Similar Tokens for Multi-View 3D Reconstruction
Zhenwei Zhu, Liying Yang, Ning Li, Chaohao Jiang, Yanyan Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18226-18235
In recent years, many video tasks have achieved breakthroughs by utilizing the vision transformer and establishing spatial-temporal decoupling for feature extraction. Although multi-view 3D reconstruction also faces multiple images as input, it cannot immediately inherit their success due to completely ambiguous associations between unstructured views. There is not usable prior relationship, which

 is similar to the temporally-coherence property in a video. To solve this probl
em, we propose a novel transformer network for Unstructured Multiple Images (UMI
Former). It exploits transformer blocks for decoupled intra-view encoding and de
signed blocks for token rectification that mine the correlation between similar
tokens from different views to achieve decoupled inter-view encoding. Afterward,
 all tokens acquired from various branches are compressed into a fixed-size comp
act representation while preserving rich information for reconstruction by lever
aging the similarities between tokens. We empirically demonstrate on ShapeNet an
d confirm that our decoupled learning method is adaptable for unstructured multi
ple images. Meanwhile, the experiments also verify our model outperforms existin
g SOTA methods by a large margin. Code will be available at https://github.com/G
aryZhu1996/UMIFormer.
****
Improved Knowledge Transfer for Semi-Supervised Domain Adaptation via Trico Trai
ning Strategy
Ba Hung Ngo, Yeon Jeong Chae, Jung Eun Kwon, Jae Hyeon Park, Sung In Cho; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 19214-19223
The motivation of the semi-supervised domain adaptation (SSDA) is to train a mod
el by leveraging knowledge acquired from the plentiful labeled source combined w
ith extremely scarce labeled target data to achieve the lowest error on the unla
beled target data at the testing time. However, due to inter-domain and intra-do
main discrepancies, the improvement of classification accuracy is limited. To so
lve these, we propose the Trico-training method that utilizes a multilayer perce
ptron (MLP) classifier and two graph convolutional network (GCN) classifiers cal
led inter-view GCN and intra-view GCN classifiers. The first co-training strateg
y exploits a correlation between MLP and inter-view GCN classifiers to minimize
the inter-domain discrepancy, in which the inter-view GCN classifier provides it
s pseudo labels to teach the MLP classifier, which encourages class representati
on alignment across domains. In contrast, the MLP classifier gives feedback to t
he inter-view GCN classifier by using a new concept, 'pseudo-edge', for neighbor
's feature aggregation. Doing this increases the data structure mining ability o
f the inter-view GCN classifier; thus, the quality of generated pseudo labels is
 improved. The second co-training strategy between MLP and intra-view GCN is con
ducted in a similar way to reduce the intra-domain discrepancy by enhancing the
correlation between labeled and unlabeled target data. Due to an imbalance in cl
assification accuracy between inter-view and intra-view GCN classifiers, we prop
ose the third co-training strategy that encourages them to cooperate to address
this problem. We verify the effectiveness of the proposed method on three standa
rd SSDA benchmark datasets: Office-31, Office-Home, and DomainNet. The extended
experimental results show that our method surpasses the prior state-of-the-art a
pproaches in SSDA.
****
Locally Stylized Neural Radiance Fields
Hong-Wing Pang, Binh-Son Hua, Sai-Kit Yeung; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2023, pp. 307-316
In recent years, there has been increasing interest in applying stylization on 3
D scenes from a reference style image, in particular onto neural radiance fields
 (NeRF). While performing stylization directly on NeRF guarantees appearance con
sistency over arbitrary novel views, it is a challenging problem to guide the tr
ansfer of patterns from the style image onto different parts of the NeRF scene.
In this work, we propose a stylization framework for NeRF based on local style t
ransfer. In particular, we use a hash-grid encoding to learn the embedding of th
e appearance and geometry components, and show that the mapping defined by the h
ash table allows us to control the stylization to a certain extent. Stylization
is then achieved by optimizing the appearance branch while keeping the geometry
branch fixed. To support local style transfer, we propose a new loss function th
at utilizes a segmentation network and bipartite matching to establish region co
rrespondences between the style image and the content images obtained from volum
e rendering. Our experiments show that our method yields plausible stylization r

esults with novel view synthesis while having flexible controllability via manip
ulating and customizing the region correspondences.
****
InterFormer: Real-time Interactive Image Segmentation
You Huang, Hao Yang, Ke Sun, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Rongr
ong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2023, pp. 22301-22311
Interactive image segmentation enables annotators to efficiently perform pixel-l
evel annotation for segmentation tasks. However, the existing interactive segmen
tation pipeline suffers from inefficient computations of interactive models beca
use of the following two issues. First, annotators' later click is based on mode
ls' feedback of annotators' former click. This serial interaction is unable to u
tilize model's parallelism capabilities. Second, in each interaction step, the m
odel handles the invariant image along with the sparse variable clicks, resultin
g in a process that's highly repetitive and redundant. For efficient computation
s, we propose a method named InterFormer that follows a new pipeline to address
these issues. InterFormer extracts and preprocesses the computationally time-con
suming part i.e. image processing from the existing process. Specifically, Inter
Former employs a large vision transformer (ViT) on high-performance devices to p
reprocess images in parallel, and then uses a lightweight module called interact
ive multi-head self attention (I-MSA) for interactive segmentation. Furthermore,
 the I-MSA module's deployment on low-power devices extends the practical applic
ation of interactive segmentation. The I-MSA module utilizes the preprocessed fe
atures to efficiently response to the annotator inputs in real-time. The experim
ents on several datasets demonstrate the effectiveness of InterFormer, which out
performs previous interactive segmentation models in terms of computational effi
ciency and segmentation quality, achieve real-time high-quality interactive segm
entation on CPU-only devices. The code is available at https://github.com/YouHua
ng67/InterFormer.
****
Confidence-aware Pseudo-label Learning for Weakly Supervised Visual Grounding
Yang Liu, Jiahua Zhang, Qingchao Chen, Yuxin Peng; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 2828-2838
Visual grounding aims at localizing the target object in image which is most rel
ated to the given free-form natural language query. As labeling the position of
target object is labor-intensive, the weakly supervised methods, where only imag
e-sentence annotations are required during model training have recently received
 increasing attention. Most of the existing weakly-supervised methods first gene
rate region proposals via pre-trained object detectors and then employ either cr
oss-modal similarity score or reconstruction loss as the criteria to select prop
osal from them. However, due to the cross-modal heterogeneous gap, these method
often suffer from high confidence spurious association and model prone to error
propagation. In this paper, we propose Confidence-aware Pseudo-label Learning (C
PL) to overcome the above limitations. Specifically, we first adopt both the uni
-modal and cross-modal pre-trained models and propose conditional prompt enginee
ring to automatically generate multiple `descriptive, realistic and diverse' pse
udo language queries for each region proposal, and then establish reliable cross
-modal association for model training based on the uni-modal similarity score (b
etween pseudo and real text queries). Secondly, we propose a confidence-aware ps
eudo label verification module which reduces the amount of noise encountered in
the training process and the risk of error propagation. Experiments on five wide
ly used datasets validate the efficacy of our proposed components and demonstrat
e state-of-the-art performance.
****
Luminance-aware Color Transform for Multiple Exposure Correction
Jong-Hyeon Baek, DaeHyun Kim, Su-Min Choi, Hyo-jun Lee, Hanul Kim, Yeong Jun Koh
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2023, pp. 6156-6165
Images captured with irregular exposures inevitably present unsatisfactory visua
l effects, such as distorted hue and color tone. However, most recent studies ma

inly focus on underexposure correction, which limits their applicability to real-world scenarios where exposure levels vary. Furthermore, some works to tackle multiple exposure rely on the encoder-decoder architecture, resulting in losses of details in input images during down-sampling and up-sampling processes. With this regard, a novel correction algorithm for multiple exposure, called luminance-aware color transform (LACT), is proposed in this study. First, we reason the relative exposure condition between images to obtain luminance features based on a luminance comparison module. Next, we encode the set of transformation functions from the luminance features, which enable complex color transformations for both overexposure and underexposure images. Finally, we project the transformed representation onto RGB color space to produce exposure correction results. Extensive experiments demonstrate that the proposed LACT yields new state-of-the-arts on two multiple exposure datasets.
****
A Simple Framework for Open-Vocabulary Segmentation and Detection
Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1020-1031
In this work, we present OpenSeeD, a simple Open-vocabulary Segmentation and Detection framework that learns from different segmentation and detection datasets. To bridge the gap of vocabulary and annotation granularity, we first introduce a pretrained text encoder to encode all the visual concepts in two tasks and learn a common semantic space for them. This gives us reasonably good results compared with the counterparts trained on segmentation task only. To further reconcile them, we locate two discrepancies: i) task discrepancy -- segmentation requires extracting masks for both foreground objects and background stuff, while detection merely cares about the former; ii) data discrepancy -- box and mask annotations are with different spatial granularity, and thus not directly interchangeable. We propose a decoupled foreground/background decoding and a conditioned mask decoding to address these issues, respectively. To this end, we develop a simple encoder-decoder model encompassing all three techniques and train it jointly on COCO and Objects365. After pretraining, our model exhibits competitive or stronger zero-shot transferability for both segmentation and detection. Specifically, OpenSeeD beats the state-of-the-art method for open-vocabulary instance and panoptic segmentation across 5 datasets, and outperforms previous work for open-vocabulary detection on LVIS and ODinW under similar settings. When transferred to specific tasks, our model achieves new SoTA on panoptic segmentation on COCO and ADE20K, and instance segmentation on ADE20K and Cityscapes.
Finally, we note that OpenSeed is the first to explore the potential of joint training on segmentation and detection, and hope it can be received as a strong baseline for developing a single model for open-vocabulary segmentation and detection.
****
Alignment Before Aggregation: Trajectory Memory Retrieval Network for Video Object Segmentation
Rui Sun, Yuan Wang, Huayu Mai, Tianzhu Zhang, Feng Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1218-1228
Memory-based methods in semi-supervised video object segmentation task achieve competitive performance by performing dense matching between query and memory frames. However, most of the existing methods neglect the fact that videos carry rich temporal information yet redundant spatial information. In this case, direct pixel-level global matching will lead to ambiguous correspondences. In this work, we reconcile the inherent tension of spatial and temporal information to retrieve memory frame information along the object trajectory, and propose a novel and coherent Trajectory Memory Retrieval Network (TMRN) to equip with the trajectory information, including a spatial alignment module and a temporal aggregation module. The proposed TMRN enjoys several merits. First, TMRN
is empowered to characterize the temporal correspondence which is in line with the nature of video in a data-driven manner. Second, we elegantly customize the spatial alignment module by coupling SVD initialization with agent-level correla

tion for representative agent construction and rectifying false matches caused by direct pairwise pixel-level correlation, respectively. Extensive experimental results on challenging benchmarks including DAVIS 2017 validation / test and Youtube-VOS 2018 / 2019 demonstrate that our TMRN, as a general plugin module, achieves consnsistent improvements over several leading methods.
****
UATVR: Uncertainty-Adaptive Text-Video Retrieval
Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13723-13733
With the explosive growth of web videos and emerging large-scale vision-language pre-training models, e.g., CLIP, retrieving videos of interest with text instructions has attracted increasing attention. A common practice is to transfer text-video pairs to the same embedding space and craft cross-modal interactions with certain entities in specific granularities for semantic correspondence. Unfortunately, the intrinsic uncertainties of optimal entity combinations in appropriate granularities for cross-modal queries are understudied, which is especially critical for modalities with hierarchical semantics, e.g., video, text, etc. In this paper, we propose an Uncertainty-Adaptive Text-Video Retrieval approach, termed UATVR, which models each look-up as a distribution matching procedure. Concretely, we add additional learnable tokens in the encoders to adaptively aggregate multi-grained semantics for flexible high-level reasoning. In the refined embedding space, we represent text-video pairs as probabilistic distributions where prototypes are sampled for matching evaluation. Comprehensive experiments on four benchmarks justify the superiority of our UATVR, which achieves new state-of-the-art results on MSR-VTT (50.8%), VATEX (64.5%), MSVD (49.7%), and DiDeMo (45.8%). The code is available at https://github.com/bofang98/UATVR.
****
Deep Directly-Trained Spiking Neural Networks for Object Detection
Qiaoyi Su, Yuhong Chou, Yifan Hu, Jianing Li, Shijie Mei, Ziyang Zhang, Guoqi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6555-6565
Spiking neural networks (SNNs) are brain-inspired energy-efficient models that encode information in spatiotemporal dynamics. Recently, deep SNNs trained directly have shown great success in achieving high performance on classification tasks with very few time steps. However, how to design a directly-trained SNN for the regression task of object detection still remains a challenging problem. To address this problem, we propose EMS-YOLO, a novel directly-trained SNN framework for object detection, which is the first trial to train a deep SNN with surrogate gradients for object detection rather than ANN-SNN conversion strategies. Specifically, we design a full-spike residual block, EMS-ResNet, which can effectively extend the depth of the directly-trained SNN with low power consumption. Furthermore, we theoretically analyze and prove the EMS-ResNet could avoid gradient vanishing or exploding. The results demonstrate that our approach outperforms the state-of-the-art ANN-SNN conversion methods (at least 500 time steps) in extremely fewer time steps (only 4 time steps). It is shown that our model could achieve comparable performance to the ANN with the same architecture while consuming 5.83x less energy on the frame-based COCO Dataset and the event-based Gen1 Dataset. Our code is available in https://github.com/BICLab/EMS-YOLO.
****
Online Prototype Learning for Online Continual Learning
Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, Hongming Shan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18764-18774
Online continual learning (CL) studies the problem of learning continuously from a single-pass data stream while adapting to new data and mitigating catastrophic forgetting. Recently, by storing a small subset of old data, replay-based methods have shown promising performance. Unlike previous methods that focus on sample storage or knowledge distillation against catastrophic forgetting, this paper

aims to understand why the online learning models fail to generalize well from a new perspective of shortcut learning. We identify shortcut learning as the key limiting factor for online CL, where the learned features may be biased, not generalizable to new tasks, and may have an adverse impact on knowledge distillation. To tackle this issue, we present the online prototype learning (OnPro) framework for online CL. First, we propose online prototype equilibrium to learn representative features against shortcut learning and discriminative features to avoid class confusion, ultimately achieving an equilibrium status that separates all seen classes well while learning new classes. Second, with the feedback of online prototypes, we devise a novel adaptive prototypical feedback mechanism to sense the classes that are easily misclassified and then enhance their boundaries. Extensive experimental results on widely-used benchmark datasets demonstrate the superior performance of OnPro over the state-of-the-art baseline methods. Source code is available at https://github.com/weilllllls/OnPro.
****
Robust e-NeRF: NeRF from Sparse & Noisy Events under Non-Uniform Motion
Weng Fei Low, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18335-18346
Event cameras offer many advantages over standard cameras due to their distinctive principle of operation: low power, low latency, high temporal resolution and high dynamic range. Nonetheless, the success of many downstream visual applications also hinges on an efficient and effective scene representation, where Neural Radiance Field (NeRF) is seen as the leading candidate. Such promise and potential of event cameras and NeRF inspired recent works to investigate on the reconstruction of NeRF from moving event cameras. However, these works are mainly limited in terms of the dependence on dense and low-noise event streams, as well as generalization to arbitrary contrast threshold values and camera speed profiles. In this work, we propose Robust e-NeRF, a novel method to directly and robustly reconstruct NeRFs from moving event cameras under various real-world conditions, especially from sparse and noisy events generated under non-uniform motion. It consists of two key components: a realistic event generation model that accounts for various intrinsic parameters (e.g. time-independent, asymmetric threshold and refractory period) and non-idealities (e.g. pixel-to-pixel threshold variation), as well as a complementary pair of normalized reconstruction losses that can effectively generalize to arbitrary speed profiles and intrinsic parameter values without such prior knowledge. Experiments on real and novel realistically simulated sequences verify our effectiveness. Our code, synthetic dataset and improved event simulator are public.
****
ActorsNeRF: Animatable Few-shot Human Rendering with Generalizable NeRFs
Jiteng Mu, Shen Sang, Nuno Vasconcelos, Xiaolong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18391-18401
While NeRF-based human representations have shown impressive novel view synthesis results, most methods still rely on a large number of images / views for training. In this work, we propose a novel animatable NeRF called ActorsNeRF. It is first pre-trained on diverse human subjects, and then adapted with few-shot monocular video frames for a new actor with unseen poses. Building on previous generalizable NeRFs with parameter sharing using a ConvNet encoder, ActorsNeRF further adopts two human priors to capture the large human appearance, shape, and pose variations. Specifically, in the encoded feature space, we will first align different human subjects in a category-level canonical space, and then align the same human from different frames in an instance-level canonical space for rendering. We quantitatively and qualitatively demonstrate that ActorsNeRF significantly outperforms the existing state-of-the-art on few-shot generalization to new people and poses on multiple datasets.
****
SALAD: Part-Level Latent Diffusion for 3D Shape Generation and Manipulation
Juil Koo, Seungwoo Yoo, Minh Hieu Nguyen, Minhyuk Sung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14441-14451
We present a cascaded diffusion model based on a part-level implicit 3D represen

tation. Our model achieves state-of-the-art generation quality and also enables part-level shape editing and manipulation without any additional training in conditional setup. Diffusion models have demonstrated impressive capabilities in data generation as well as zero-shot completion and editing via a guided reverse process. Recent research on 3D diffusion models has focused on improving their generation capabilities with various data representations, while the absence of structural information has limited their capability in completion and editing tasks. We thus propose our novel diffusion model using a part-level implicit representation. To effectively learn diffusion with high-dimensional embedding vectors of parts, we propose a cascaded framework, learning diffusion first on a low-dimensional subspace encoding extrinsic parameters of parts and then on the other high-dimensional subspace encoding intrinsic attributes. In the experiments, we demonstrate the outperformance of our method compared with the previous ones both in generation and part-level completion and manipulation tasks.
****
COMPASS: High-Efficiency Deep Image Compression with Arbitrary-scale Spatial Scalability
Jongmin Park, Jooyoung Lee, Munchurl Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12826-12835
Recently, neural network (NN)-based image compression studies have actively been made and has shown impressive performance in comparison to traditional methods. However, most of the works have focused on non-scalable image compression (single-layer coding) while spatially scalable image compression has drawn less attention although it has many applications. In this paper, we propose a novel NN-based spatially scalable image compression method, called COMPASS, which supports arbitrary-scale spatial scalability. Our proposed COMPASS has a very flexible structure where the number of layers and their respective scale factors can be arbitrarily determined during inference. To reduce the spatial redundancy between adjacent layers for arbitrary scale factors, our COMPASS adopts an inter-layer arbitrary scale prediction method, called LIFF, based on implicit neural representation. We propose a combined RD loss function to effectively train multiple layers. Experimental results show that our COMPASS achieves BD-rate gain of -58.33% and -47.17% at maximum compared to SHVC and the state-of-the-art NN-based spatially scalable image compression method, respectively, for various combinations of scale factors. Our COMPASS also shows comparable or even better coding efficiency than the single-layer coding for various scale factors.
****
Masked Autoencoders Are Stronger Knowledge Distillers
Shanshan Lao, Guanglu Song, Boxiao Liu, Yu Liu, Yujiu Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6384-6393
Knowledge distillation (KD) has shown great success in improving student's performance by mimicking the intermediate output of the high-capacity teacher in fine-grained visual tasks, e.g. object detection. This paper proposes a technique called Masked Knowledge Distillation (MKD) that enhances this process using a masked autoencoding scheme. In MKD, random patches of the input image are masked, and the corresponding missing feature is recovered by forcing it to imitate the output of the teacher. MKD is based on two core designs. First, using the student as the encoder, we develop an adaptive decoder architecture, which includes a spatial alignment module that operates on the multi-scale features in the feature pyramid network (FPN), a simple decoder, and a spatial recovery module that mimics the teacher's output from the latent representation and mask tokens. Second, we introduce the masked convolution in each convolution block to keep the masked patches unaffected by others. By coupling these two designs, we can further improve the completeness and effectiveness of teacher knowledge learning. We conduct extensive experiments on different architectures with object detection and semantic segmentation. The results show that all the students can achieve further improvements compared to the conventional KD. Notably, we establish the new state-of-the-art results by boosting RetinaNet ResNet-18, and ResNet-50 from 33.4 to 37.5 mAP, and 37.4 to 41.5 mAP, respectively.
****

Score-Based Diffusion Models as Principled Priors for Inverse Imaging
Berthy T. Feng, Jamie Smith, Michael Rubinstein, Huiwen Chang, Katherine L. Bouman, William T. Freeman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10520-10531
Priors are essential for reconstructing images from noisy and/or incomplete measurements. The choice of the prior determines both the quality and uncertainty of recovered images. We propose turning score-based diffusion models into principled image priors ("score-based priors") for analyzing a posterior of images given measurements. Previously, probabilistic priors were limited to handcrafted regularizers and simple distributions. In this work, we empirically validate the theoretically-proven probability function of a score-based diffusion model. We show how to sample from resulting posteriors by using this probability function for variational inference. Our results, including experiments on denoising, deblurring, and interferometric imaging, suggest that score-based priors enable principled inference with a sophisticated, data-driven image prior.
****
Multiscale Structure Guided Diffusion for Image Deblurring
Mengwei Ren, Mauricio Delbracio, Hossein Talebi, Guido Gerig, Peyman Milanfar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10721-10733
Diffusion Probabilistic Models (DPMs) have recently been employed for image deblurring, formulated as an image-conditioned generation process that maps Gaussian noise to the high-quality image, conditioned on the blurry input. Image-conditioned DPMs (icDPMs) have shown more realistic results than regression-based methods when trained on pairwise in-domain data. However, their robustness in restoring images is unclear when presented with out-of-domain images as they do not impose specific degradation models or intermediate constraints. To this end, we introduce a simple yet effective multiscale structure guidance as an implicit bias that informs the icDPM about the coarse structure of the sharp image at the intermediate layers. This guided formulation leads to a significant improvement of the deblurring results, particularly on unseen domain. The guidance is extracted from the latent space of a regression network trained to predict the clean-sharp target at multiple lower resolutions, thus maintaining the most salient sharp structures. With both the blurry input and multiscale guidance, the icDPM model can better understand the blur and recover the clean image. We evaluate a single-dataset trained model on diverse datasets and demonstrate more robust deblurring results with fewer artifacts on unseen data. Our method outperforms existing baselines, achieving state-of-the-art perceptual quality while keeping competitive distortion metrics.
****
Multiple Planar Object Tracking
Zhicheng Zhang, Shengzhe Liu, Jufeng Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23460-23470
Tracking both location and pose of multiple planar objects (MPOT) is of great significance to numerous real-world applications. The greater degree-of-freedom of planar objects compared with common objects makes MPOT far more challenging than well-studied object tracking, especially when occlusion occurs. To address this challenging task, we are inspired by amodal perception that humans jointly track visible and invisible parts of the target, and propose a tracking framework that unifies appearance perception and occlusion reasoning. Specifically, we present a dual-branch network to track the visible part of planar objects, including vertexes and mask. Then, we develop an occlusion area localization strategy to infer the invisible part, i.e., the occluded region, followed by a two-stream attention network finally refining the prediction. To alleviate the lack of data in this field, we build the first large-scale benchmark dataset, namely MPOT-3K. It consists of 3,717 planar objects from 356 videos and contains 148,896 frames together with 687,417 annotations. The collected planar objects have 9 motion patterns and the videos are shot in 6 types of indoor and outdoor scenes. Extensive experiments demonstrate the superiority of our proposed method on the newly developed MPOT-3K as well as other two popular single planar object tracking datas

ets. The code and MPOT-3K dataset are released on https://zzcheng.top/MPOT.
****
CheckerPose: Progressive Dense Keypoint Localization for Object Pose Estimation with Graph Neural Network
Ruyi Lian, Haibin Ling; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14022-14033
Estimating the 6-DoF pose of a rigid object from a single RGB image is a crucial yet challenging task. Recent studies have shown the great potential of dense correspondence-based solutions, yet improvements are still needed to reach practical deployment. In this paper, we propose a novel pose estimation algorithm named CheckerPose, which improves on three main aspects. Firstly, CheckerPose densely samples 3D keypoints from the surface of the 3D object and finds their 2D correspondences progressively in the 2D image. Compared to previous solutions that conduct dense sampling in the image space, our strategy enables the correspondence searching in a 2D grid (i.e., pixel coordinate). Secondly, for our 3D-to-2D correspondence, we design a compact binary code representation for 2D image locations. This representation not only allows for progressive correspondence refinement but also converts the correspondence regression to a more efficient classification problem. Thirdly, we adopt a graph neural network to explicitly model the interactions among the sampled 3D keypoints, further boosting the reliability and accuracy of the correspondences. Together, these novel components make CheckerPose a strong pose estimation algorithm. When evaluated on the popular Linemod, Linemod-O, and YCB-V object pose estimation benchmarks, CheckerPose clearly boosts the accuracy of correspondence-based methods and achieves state-of-the-art performances. Code is available at https://github.com/RuyiLian/CheckerPose.
****
ASIC: Aligning Sparse in-the-wild Image Collections
Kamal Gupta, Varun Jampani, Carlos Esteves, Abhinav Shrivastava, Ameesh Makadia, Noah Snavely, Abhishek Kar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4134-4145
We present a method for joint alignment of sparse in-the-wild image collections of an object category. Most prior works assume either ground-truth keypoint annotations or a large dataset of images of a single object category. However, neither of the above assumptions hold true for the long-tail of the objects present in the world. We present a self-supervised technique that directly optimizes on a sparse collection of images of a particular object/object category to obtain consistent dense correspondences across the collection. We use pairwise nearest neighbors obtained from deep features of a pre-trained vision transformer (ViT) model as noisy and sparse keypoint matches and make them dense and accurate matches by optimizing a neural network that jointly maps the image collection into a learned canonical grid. Experiments on CUB, SPair-71k and PF-Willow benchmarks demonstrate that our method can produce globally consistent and higher quality correspondences across the image collection when compared to existing self-supervised methods. Code and other material will be made available at https://kampta.github.io/asic.
****
Residual Pattern Learning for Pixel-Wise Out-of-Distribution Detection in Semantic Segmentation
Yuyuan Liu, Choubo Ding, Yu Tian, Guansong Pang, Vasileios Belagiannis, Ian Reid, Gustavo Carneiro; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1151-1161
Semantic segmentation models classify pixels into a set of known ("in-distribution") visual classes. When deployed in an open world, the reliability of these models depends on their ability to not only classify in-distribution pixels but also to detect out-of-distribution (OoD) pixels.

Historically, the poor OoD detection performance of these models has motivated the design of methods based on model re-training using synthetic training images that include OoD visual objects.

Although successful, these re-trained methods have two issues: 1) their in-distribution segmentation accuracy may drop during re-training, and 2) their OoD det

ection accuracy does not generalise well to new contexts (e.g., country surround ings) outside the training set (e.g., city surroundings).
In this paper, we mitigate these issues with: (i) a new residual pattern learni ng (RPL) module that assists the segmentation model to detect OoD pixels with mi nimal deterioration to the inlier segmentation performance; and (ii) a novel con text-robust contrastive learning (CoroCL) that enforces RPL to robustly detect O oD pixels in various contexts. Our approach improves by around 10% FPR and 7% Au PRC the previous state-of-the-art in Fishyscapes, Segment-Me-If-You-Can, and Roa dAnomaly datasets. Code will be available.
****
Hierarchical Visual Primitive Experts for Compositional Zero-Shot Learning
Hanjae Kim, Jiyoung Lee, Seongheon Park, Kwanghoon Sohn; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2023, pp. 5675-5685
Compositional zero-shot learning (CZSL) aims to recognize unseen compositions wi th prior knowledge of known primitives (attribute and object). Previous works fo r CZSL often suffer from grasping the contextuality between attribute and object , as well as the discriminability of visual features, and the long-tailed distri bution of real-world compositional data. We propose a simple and scalable framew ork called Composition Transformer (CoT) to address these issues. CoT employs ob ject and attribute experts in distinctive manners to generate representative emb eddings, using the visual network hierarchically. The object expert extracts rep resentative object embeddings from the final layer in a bottom-up manner, while the attribute expert makes attribute embeddings in a top-down manner with a prop osed object-guided attention module that models contextuality explicitly. To rem edy biased prediction caused by imbalanced data distribution, we develop a simpl e minority attribute augmentation (MAA) that synthesizes virtual samples by mixi ng two images and oversampling minority attribute classes. Our method achieves S oTA performance on several benchmarks, including MIT-States, C-GQA, and VAW-CZSL . We also demonstrate the effectiveness of CoT in improving visual discriminatio n and addressing the model bias from the imbalanced data distribution. The code is available at https://github.com/HanjaeKim98/CoT.
****
Event Camera Data Pre-training
Yan Yang, Liyuan Pan, Liu Liu; Proceedings of the IEEE/CVF International Confere nce on Computer Vision (ICCV), 2023, pp. 10699-10709
This paper proposes a pre-trained neural network for handling event camera data. Our model is a self-supervised learning framework, and uses paired event camera data and natural RGB images for training. Our method contains three modules con nected in a sequence: i) a family of event data augmentations, generating meanin gful event images for self-supervised training; ii) a conditional masking strate gy to sample informative event patches from event images, encouraging our model to capture the spatial layout of a scene and accelerating training; iii) a contr astive learning approach, enforcing the similarity of embeddings between matchin g event images, and between paired event and RGB images. An embedding projection loss is proposed to avoid the model collapse when enforcing the event image emb edding similarities. A probability distribution alignment loss is proposed to en courage the event image to be consistent with its paired RGB image in the featur e space. Transfer learning performance on downstream tasks shows the superiority of our method over state-of-the-art methods. For example, we achieve top-1 accu racy at 64.83% on the N-ImageNet dataset. Our code is available at https://githu b.com/Yan98/Event-Camera-Data-Pre-training.
****
Segment Every Reference Object in Spatial and Temporal Spaces
Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2538 -2550
The reference-based object segmentation tasks, namely referring image segmentati on (RIS), referring video object segmentation (RVOS), and video object segmentat ion (VOS), aim to segment a specific object by utilizing either
language or annotated masks as references. Despite significant progress in each

respective field, current methods are task-specifically designed and developed in different directions, which hinders the activation of multi-task capabilities for these tasks. In this work, we end the current fragmented situation and propose UniRef to unify the three reference-based object segmentation tasks with a single architecture. At the heart of our approach is the multiway-fusion for handling different task with respect to their specified references. And a unified Transformer architecture is then adopted for performing instance-level segmentation. With the unified designs, UniRef can be jointly trained on a broad range of benchmarks and can flexibly perform multiple tasks at runtime by specifying the corresponding references. We evaluate the jointly trained network on various benchmarks. Extensive experimental results indicate that our proposed UniRef achieves state-of-the-art performance on RIS and RVOS, and performs competitively on VOS with a single network.
****
Unified Out-Of-Distribution Detection: A Model-Specific Perspective
Reza Averly, Wei-Lun Chao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1453-1463
Out-of-distribution (OOD) detection aims to identify test examples that do not belong to the training distribution and are thus unlikely to be predicted reliably. Despite a plethora of existing works, most of them focused only on the scenario where OOD examples come from semantic shift (e.g., unseen categories), ignoring other possible causes (e.g., covariate shift). In this paper, we present a novel, unifying framework to study OOD detection in a broader scope. Instead of detecting OOD examples from a particular cause, we propose to detect examples that a deployed machine learning model (e.g., an image classifier) is unable to predict correctly. That is, whether a test example should be detected and rejected or not is "model-specific". We show that this framework unifies the detection of OOD examples caused by semantic shift and covariate shift, and closely addresses the concern of applying a machine learning model to uncontrolled environments. We provide an extensive analysis that involves a variety of models (e.g., different architectures and training strategies), sources of OOD examples, and OOD detection approaches, and reveal several insights into improving and understanding OOD detection in uncontrolled environments.
****
One-shot Implicit Animatable Avatars with Model-based Priors
Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, Deng Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8974-8985
Existing neural rendering methods for creating human avatars typically either require dense input signals such as video or multi-view images, or leverage a learned prior from large-scale specific 3D human datasets such that reconstruction can be performed with sparse-view inputs. Most of these methods fail to achieve realistic reconstruction when only a single image is available. To enable the data-efficient creation of realistic animatable 3D humans, we propose ELICIT, a novel method for learning human-specific neural radiance fields from a single image. Inspired by the fact that humans can effortlessly estimate the body geometry and imagine full-body clothing from a single image, we leverage two priors in ELICIT: 3D geometry prior and visual semantic prior. Specifically, ELICIT utilizes the 3D body shape geometry prior from a skinned vertex-based template model (i.e., SMPL) and implements the visual clothing semantic prior with the CLIP-based pretrained models. Both priors are used to jointly guide the optimization for creating plausible content in the invisible areas. Taking advantage of the CLIP models, ELICIT can use text descriptions to generate text-conditioned unseen regions. In order to further improve visual details, we propose a segmentation-based sampling strategy that locally refines different parts of the avatar. Comprehensive evaluations on multiple popular benchmarks, including ZJU-MoCAP, Human3.6M, and DeepFashion, show that ELICIT has outperformed strong baseline methods of avatar creation when only a single image is available. The code is public for research purposes at https://huangyangyi.github.io/ELICIT/.
****

Unsupervised Feature Representation Learning for Domain-generalized Cross-domain Image Retrieval
Conghui Hu, Can Zhang, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11016-11025
Cross-domain image retrieval has been extensively studied due to its high practical value. In recently proposed unsupervised cross-domain image retrieval methods, efforts are taken to break the data annotation barrier. However, applicability of the model is still confined to domains seen during training. This limitation motivates us to present the first attempt at domain-generalized unsupervised cross-domain image retrieval (DG-UCDIR) aiming at facilitating image retrieval between any two unseen domains in an unsupervised way. To improve domain generalizability of the model, we thus propose a new two-stage domain augmentation technique for diversified training data generation. DG-UCDIR also shares all the challenges present in the unsupervised cross-domain image retrieval, where domain-agnostic and semantic-aware feature representations are supposed to be learned without external supervision. To accomplish this, we introduce a novel cross-domain contrastive learning strategy by utilizing phase image as a proxy to mitigate the domain gap. Extensive experiments are carried out using PACS and DomainNet dataset, and consistently illustrate the superior performance of our framework compared to existing state-of-the-art methods. Our source code is available at https : //github.com/conghui1002/DG-UCDIR.
****
RankMatch: Fostering Confidence and Consistency in Learning with Noisy Labels
Ziyi Zhang, Weikai Chen, Chaowei Fang, Zhen Li, Lechao Chen, Liang Lin, Guanbin Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1644-1654
Learning with noisy labels (LNL) is one of the most important and challenging problems in weakly-supervised learning. Recent advances adopt the sample selection strategy to mitigate the interference of noisy labels and use small-loss criteria to select clean samples. However, the one-dimensional loss is an over-simplified metric that fails to accommodate the complex feature landscape of various samples, and, hence, is prone to introduce classification errors during sample selection. In this paper, we propose RankMatch, a novel LNL framework that investigates additional dimensions of confidence and consistency in order to combat noisy labels. Confidence-wise, we propose a novel sample selection strategy based on confidence representation voting instead of the widely-used small-loss criterion. This new strategy is capable of increasing sample selection quantity without sacrificing labeling accuracy. Consistency-wise, instead of the widely adopted feature distance metric for measuring the consistency of inner-class samples, we advocate that the rank of principal features is a much more robust indicator. Based on this metric, we propose rank contrastive loss, which strengthens the consistency of similar samples regardless of their labels and facilitates feature representation learning. Experimental results on noisy versions of CIFAR-10, CIFAR-100, Clothing1M, and WebVision have validated the superiority of our approach over existing state-of-the-art methods.
****
Dec-Adapter: Exploring Efficient Decoder-Side Adapter for Bridging Screen Content and Natural Image Compression
Sheng Shen, Huanjing Yue, Jingyu Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12887-12896
Natural image compression has been greatly improved in the deep learning era. However, the compression performance will be heavily degraded if the pretrained encoder is directly applied on screen content image compression. Meanwhile, we observe that parameter-efficient trans-fer learning (PETL) methods have shown great adaptation ability in high-level vision tasks. Therefore, we propose a Dec-Adapter, a pioneering entropy-efficient transfer learning module for the decoder to bridge natural image and screen content compression. The adapter's parameters are learned during encoding and transmitted to the decoder for image-adaptive decoding. Our Dec-Adapter is lightweight, domain-transferable, and architecture-agnostic with generalized performance in bridging the two domains. Experiments demon

strate that our method outperforms all existing methods by a large margin in terms of BD-rate performance on screen content image compression. Specifically, our method achieves over 2 dB gain compared with the baseline when transferred to screen content image com-pression.
****
MixReorg: Cross-Modal Mixed Patch Reorganization is a Good Mask Learner for Open-World Semantic Segmentation
Kaixin Cai, Pengzhen Ren, Yi Zhu, Hang Xu, Jianzhuang Liu, Changlin Li, Guangrun Wang, Xiaodan Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1196-1205
Recently, semantic segmentation models trained with image-level text supervision have shown promising results in challenging open-world scenarios. However, these models still face difficulties in learning fine-grained semantic alignment at the pixel level and predicting accurate object masks. To address this issue, we propose MixReorg, a novel and straightforward pre-training paradigm for semantic segmentation that enhances a model's ability to reorganize patches mixed across images, exploring both local visual relevance and global semantic coherence. Our approach involves generating fine-grained patch-text pairs data by mixing image patches while preserving the correspondence between patches and text. The model is then trained to minimize the segmentation loss of the mixed images and the two contrastive losses of the original and restored features. With MixReorg as a mask learner, conventional text-supervised semantic segmentation models can achieve highly generalizable pixel-semantic alignment ability, which is crucial for open-world segmentation. After training with large-scale image-text data, MixReorg models can be applied directly to segment visual objects of arbitrary categories, without the need for further fine-tuning. Our proposed framework demonstrates strong performance on popular zero-shot semantic segmentation benchmarks, outperforming GroupViT by significant margins of 5.0%, 6.2%, 2.5%, and 3.4% mIoU on PASCAL VOC2012, PASCAL Context, MS COCO, and ADE20K, respectively.
****
Preface: A Data-driven Volumetric Prior for Few-shot Ultra High-resolution Face Synthesis
Marcel C. Bühler, Kripasindhu Sarkar, Tanmay Shah, Gengyan Li, Daoye Wang, Leonhard Helminger, Sergio Orts-Escolano, Dmitry Lagun, Otmar Hilliges, Thabo Beeler, Abhimitra Meka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3402-3413
NeRFs have enabled highly realistic synthesis of human faces including complex appearance and reflectance effects of hair and skin. These methods typically require a large number of multi-view input images, making the process hardware intensive and cumbersome, limiting applicability to unconstrained settings. We propose a novel volumetric human face prior that enables the synthesis of ultra high-resolution novel views of subjects that are not part of the prior's training distribution. This prior model consists of an identity-conditioned NeRF, trained on a dataset of low-resolution multi-view images of diverse humans with known camera calibration. A simple sparse landmark-based 3D alignment of the training dataset allows our model to learn a smooth latent space of geometry and appearance despite a limited number of training identities. A high-quality volumetric representation of a novel subject can be obtained by model fitting to 2 or 3 camera views of arbitrary resolution. Importantly, our method requires as few as two views of casually captured images as input at inference time.
****
Label-Guided Knowledge Distillation for Continual Semantic Segmentation on 2D Images and 3D Point Clouds
Ze Yang, Ruibo Li, Evan Ling, Chi Zhang, Yiming Wang, Dezhao Huang, Keng Teck Ma, Minhoe Hur, Guosheng Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18601-18612
Continual semantic segmentation (CSS) aims to extend an existing model to tackle unseen tasks while retaining its old knowledge. Naively fine-tuning the old model on new data leads to catastrophic forgetting. A common solution is knowledge distillation (KD), where the output distribution of the new model is regularized

to be similar to that of the old model. However, in CSS, this is challenging because of the background shift issue. Existing KD-based CSS methods continue to suffer from confusion between the background and novel classes since they fail to establish a reliable class correspondence for distillation. To address this issue, we propose a new label-guided knowledge distillation (LGKD) loss, where the old model output is expanded and transplanted (with the guidance of the ground truth label) to form a semantically appropriate class correspondence with the new model output. Consequently, the useful knowledge from the old model can be effectively distilled into the new model without causing confusion. We conduct extensive experiments on two prevailing CSS benchmarks, Pascal-VOC and ADE20K, where our LGKD significantly boosts the performance of three competing methods, especially on novel mIoU by up to +76%, setting new state-of-the-art. Finally, to further demonstrate its generalization ability, we introduce the first CSS benchmark for 3D point cloud based on ScanNet, along with several re-implemented baselines for comparison. Experiments show that LGKD is versatile in both 2D and 3D modalities without requiring ad hoc design. Codes are available at https://github.com/Ze-Yang/LGKD.
****

Under-Display Camera Image Restoration with Scattering Effect

Binbin Song, Xiangyu Chen, Shuning Xu, Jiantao Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12580-12589

The under-display camera (UDC) provides consumers with a full-screen visual experience without any obstruction due to notches or punched holes. However, the semi-transparent nature of the display inevitably introduces the severe degradation into UDC images. In this work, we address the UDC image restoration problem with the specific consideration of the scattering effect caused by the display. We explicitly model the scattering effect by treating the display as a piece of homogeneous scattering medium. With the physical model of the scattering effect, we improve the image formation pipeline for the image synthesis to construct a realistic UDC dataset with ground truths. To suppress the scattering effect for the eventual UDC image recovery, a two-branch restoration network is designed. More specifically, the scattering branch leverages global modeling capabilities of the channel-wise self-attention to estimate parameters of the scattering effect from degraded images. While the image branch exploits the local representation advantage of CNN to recover clear scenes, implicitly guided by the scattering branch. Extensive experiments are conducted on both real-world and synthesized data, demonstrating the superiority of the proposed method over the state-of-the-art UDC restoration techniques. The source code and dataset are available at https://github.com/NamecantbeNULL/SRUDC.
****

PRANC: Pseudo RAndom Networks for Compacting Deep Models

Parsa Nooralinejad, Ali Abbasi, Soroush Abbasi Koohpayegani, Kossar Pourahmadi Meibodi, Rana Muhammad Shahroz Khan, Soheil Kolouri, Hamed Pirsiavash; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17021-17031

We demonstrate that a deep model can be reparametrized as a linear combination of several randomly initialized and frozen deep models in the weight space. During training, we seek local minima that reside within the subspace spanned by these random models (i.e., `basis' networks). Our framework, PRANC, enables significant compaction of a deep model. The model can be reconstructed using a single scalar `seed,' employed to generate the pseudo-random `basis' networks, together with the learned linear mixture coefficients. In practical applications, PRANC addresses the challenge of efficiently storing and communicating deep models, a common bottleneck in several scenarios, including multi-agent learning, continual learners, federated systems, and edge devices, among others. In this study, we employ PRANC to condense image classification models and compress images by compacting their associated implicit neural networks. PRANC outperforms baselines with a large margin on image classification when compressing a deep model almost 100 times. Moreover, we show that PRANC enables memory-efficient inference by generating layer-wise weights on the fly. The source code of PRANC is here: https://

github.com/UCDvision/PRANC
****
ICICLE: Interpretable Class Incremental Continual Learning
Dawid Rymarczyk, Joost van de Weijer, Bartosz Zieli■ski, Bartlomiej Twardowski; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1887-1898
Continual learning enables incremental learning of new tasks without forgetting those previously learned, resulting in positive knowledge transfer that can enhance performance on both new and old tasks. However, continual learning poses new challenges for interpretability, as the rationale behind model predictions may change over time, leading to interpretability concept drift.
We address this problem by proposing Interpretable Class-InCremental LEarning (ICICLE), an exemplar-free approach that adopts a prototypical part-based approach. It consists of three crucial novelties: interpretability regularization that distills previously learned concepts while preserving user-friendly positive reasoning; proximity-based prototype initialization strategy dedicated to the fine-grained setting; and task-recency bias compensation devoted to prototypical parts.
Our experimental results demonstrate that ICICLE reduces the interpretability concept drift and outperforms the existing exemplar-free methods of common class-incremental learning when applied to concept-based models.
****
Clutter Detection and Removal in 3D Scenes with View-Consistent Inpainting
Fangyin Wei, Thomas Funkhouser, Szymon Rusinkiewicz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18131-18141
Removing clutter from scenes is essential in many applications, ranging from privacy-concerned content filtering to data augmentation. In this work, we present an automatic system that removes clutter from 3D scenes and inpaints with coherent geometry and texture. We propose techniques for its two key components: 3D segmentation based on shared properties and 3D inpainting, both of which are important problems. We define 3D scene clutter as frequently-moving objects (e.g. clothes or chairs that are typically moved within a few days). The definition of 3D scene clutter (frequently-moving objects) is not well captured by commonly-studied object categories in computer vision. To tackle the lack of well-defined clutter annotations, we group noisy fine-grained labels, leverage virtual rendering, and impose an instance-level area-sensitive loss. Once clutter is removed, we inpaint geometry and texture in the resulting holes by merging inpainted RGB-D images. This requires novel voting and pruning strategies that guarantee multi-view consistency across individually inpainted images for mesh reconstruction. Experiments on ScanNet and Matterport3D dataset show that our method outperforms baselines for clutter segmentation and 3D inpainting, both visually and quantitatively. Project page: https://weify627.github.io/clutter/.
****
PointCLIP V2: Prompting CLIP and GPT for Powerful 3D Open-world Learning
Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, Peng Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2639-2650
Large-scale pre-trained models have shown promising open-world performance for both vision and language tasks. However, their transferred capacity on 3D point clouds is still limited and only constrained to the classification task. In this paper, we first collaborate CLIP and GPT to be a unified 3D open-world learner, named as PointCLIP V2, which fully unleashes their potential for zero-shot 3D classification, segmentation, and detection. To better align 3D data with the pre-trained language knowledge, PointCLIP V2 contains two key designs. For the visual end, we prompt CLIP via a shape projection module to generate more realistic depth maps, narrowing the domain gap between projected point clouds with natural images. For the textual end, we prompt the GPT model to generate 3D-specific text as the input of CLIP's textual encoder. Without any training in 3D domains, our approach significantly surpasses PointCLIP by +42.90%, +40.44%, and +28.75% accuracy on three datasets for zero-shot 3D classification. On top of that, V2 can

be extended to few-shot 3D classification, zero-shot 3D part segmentation, and 3D object detection in a simple manner, demonstrating our generalization ability for unified 3D open-world learning.
****
VideoFlow: Exploiting Temporal Cues for Multi-frame Optical Flow Estimation
Xiaoyu Shi, Zhaoyang Huang, Weikang Bian, Dasong Li, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, Jifeng Dai, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12469-12480
We introduce VideoFlow, a novel optical flow estimation framework for videos. In contrast to previous methods that learn to estimate optical flow from two frames, VideoFlow concurrently estimates bi-directional optical flows for multiple frames that are available in videos by sufficiently exploiting temporal cues. We first propose a TRi-frame Optical Flow (TROF) module that estimates bi-directional optical flows for the center frame in a three-frame manner. The information of the frame triplet is iteratively fused onto the center frame. To extend TROF for handling more frames, we further propose a MOtion Propagation (MOP) module that bridges multiple TROFs and propagates motion features between adjacent TROFs. With the iterative flow estimation refinement, the information fused in individual TROFs can be propagated into the whole sequence via MOP. By effectively exploiting video information, VideoFlow presents extraordinary performance, ranking 1st on all public benchmarks. On the Sintel benchmark, VideoFlow achieves 1.649 and 0.991 average end-point-error (AEPE) on the final and clean passes, a 15.1% and 7.6% error reduction from the best published results (1.943 and 1.073 from FlowFormer++). On the KITTI-2015 benchmark, VideoFlow achieves an F1-all error of 3.65%, a 19.2% error reduction from the best published result (4.52% from FlowFormer++). Code is released at https://github.com/XiaoyuShi97/VideoFlow.
****
3DMiner: Discovering Shapes from Large-Scale Unannotated Image Datasets
Ta-Ying Cheng, Matheus Gadelha, Sören Pirk, Thibault Groueix, Radomír M■ch, Andrew Markham, Niki Trigoni; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9331-9341
We present 3DMiner -- a pipeline for mining 3D shapes from challenging large-scale unannotated image datasets. Unlike other unsupervised 3D reconstruction methods, we assume that, within a large-enough dataset, there must exist images of objects with similar shapes but varying backgrounds, textures, and viewpoints. Our approach leverages the recent advances in learning self-supervised image representations to cluster images with geometrically similar shapes and find common image correspondences between them. We then exploit these correspondences to obtain rough camera estimates as initialization for bundle-adjustment. Finally, for every image cluster, we apply a progressive bundle-adjusting reconstruction method to learn a neural occupancy field representing the underlying shape. We show that this procedure is robust to several types of errors introduced in previous steps (e.g., wrong camera poses, images containing dissimilar shapes, etc.), allowing us to obtain shape and pose annotations for images in-the-wild.
When using images from Pix3D chairs, our method is capable of producing significantly better results than state-of-the-art unsupervised 3D
reconstruction techniques, both quantitatively and qualitatively. Furthermore, we show how 3DMiner can be applied to in-the-wild data by reconstructing shapes present in images from the LAION-5B dataset.
****
Identification of Systematic Errors of Image Classifiers on Rare Subgroups
Jan Hendrik Metzen, Robin Hutmacher, N. Grace Hua, Valentyn Boreiko, Dan Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5064-5073
Despite excellent average-case performance of many image classifiers, their performance can substantially deteriorate on semantically coherent subgroups of the data that were under-represented in the training data. These systematic errors can impact both fairness for demographic minority groups as well as robustness and safety under domain shift. A major challenge is to identify such subgroups with subpar performance when the subgroups are not annotated and their occurrence i

s very rare. We leverage recent advances in text-to-image models and search in the space of textual descriptions of subgroups ("prompts") for subgroups where the target model has low performance on the prompt-conditioned synthesized data. To tackle the exponentially growing number of subgroups, we employ combinatorial testing. We denote this procedure as PromptAttack as it can be interpreted as an adversarial attack in a prompt space. We study subgroup coverage and identifiability with PromptAttack in a controlled setting and find that it identifies systematic errors with high accuracy. Thereupon, we apply PromptAttack to ImageNet classifiers and identify novel systematic errors on rare subgroups.
****
Hierarchical Spatio-Temporal Representation Learning for Gait Recognition
Lei Wang, Bo Liu, Fangfang Liang, Bincheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19639-19649
Gait recognition is a biometric technique that identifies individuals by their unique walking styles, which is suitable for unconstrained environments and has a wide range of applications. While current methods focus on exploiting body part-based representations, they often neglect the hierarchical dependencies between local motion patterns. In this paper, we propose a hierarchical spatio-temporal representation learning (HSTL) framework for extracting gait features from coarse to fine. Our framework starts with a hierarchical clustering analysis to recover multi-level body structures from the whole body to local details. Next, an adaptive region-based motion extractor (ARME) is designed to learn region-independent motion features. The proposed HSTL then stacks multiple ARMEs in a top-down manner, with each ARME corresponding to a specific partition level of the hierarchy. An adaptive spatio-temporal pooling (ASTP) module is used to capture gait features at different levels of detail to perform hierarchical feature mapping. Finally, a frame-level temporal aggregation (FTA) module is employed to reduce redundant information in gait sequences through multi-scale temporal downsampling. Extensive experiments on CASIA-B, OUMVLP, GREW, and Gait3D datasets demonstrate that our method outperforms the state-of-the-art while maintaining a reasonable balance between model accuracy and complexity. Code is available at: https://github.com/gudaochangsheng/HSTL.
****
Order-Prompted Tag Sequence Generation for Video Tagging
Zongyang Ma, Ziqi Zhang, Yuxin Chen, Zhongang Qi, Yingmin Luo, Zekun Li, Chunfeng Yuan, Bing Li, Xiaohu Qie, Ying Shan, Weiming Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15681-15690
Video Tagging intends to infer multiple tags spanning relevant content for a given video. Typically, video tags are freely defined and uploaded by a variety of users, so they have two characteristics: abundant in quantity and disordered intra-video. It is difficult for the existing multi-label classification and generation methods to adapt directly to this task. This paper proposes a novel generative model, Order-Prompted Tag Sequence Generation (OP-TSG), according to the above characteristics. It regards video tagging as a tag sequence generation problem guided by sample-dependent order prompts. These prompts are semantically aligned with tags and enable to decouple tag generation order, making the model focus on modeling the tag dependencies. Moreover, the word-based generation strategy enables the model to generate novel tags. To verify the effectiveness and generalization of the proposed method, a Chinese video tagging benchmark CREATE-tagging, and an English image tagging benchmark Pexel-tagging are established. Extensive results show that OP-TSG is significantly superior to other methods, especially the results on rare tags improve by 3.3% and 3% over SOTA methods on CREATE-tagging and Pexel-tagging, and novel tags generated on CREATE-tagging exhibit a tag gain of 7.04%.
****
XVO: Generalized Visual Odometry via Cross-Modal Self-Training
Lei Lai, Zhongkai Shangguan, Jimuyang Zhang, Eshed Ohn-Bar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10094-10105
We propose XVO, a semi-supervised learning method for training generalized monoc

ular Visual Odometry (VO) models with robust off-the-self operation across diverse datasets and settings. In contrast to standard monocular VO approaches which often study a known calibration within a single dataset, XVO efficiently learns to recover relative pose with real-world scale from visual scene semantics, i.e., without relying on any known camera parameters. We optimize the motion estimation model via self-training from large amounts of unconstrained and heterogeneous dash camera videos available on YouTube. Our key contribution is twofold. First, we empirically demonstrate the benefits of semi-supervised training for learning a general-purpose direct VO regression network. Second, we demonstrate multi-modal supervision, including segmentation, flow, depth, and audio auxiliary prediction tasks, to facilitate generalized representations for the VO task. Specifically, we find audio prediction task to significantly enhance the semi-supervised learning process while alleviating noisy pseudo-labels, particularly in highly dynamic and out-of-domain video data. Our proposed teacher network achieves state-of-the-art performance on the commonly used KITTI benchmark despite no multi-frame optimization or knowledge of camera parameters. Combined with the proposed semi-supervised step, XVO demonstrates off-the-shelf knowledge transfer across diverse conditions on KITTI, nuScenes, and Argoverse without fine-tuning.
****
Weakly Supervised Learning of Semantic Correspondence through Cascaded Online Correspondence Refinement
Yiwen Huang, Yixuan Sun, Chenghang Lai, Qing Xu, Xiaomei Wang, Xuli Shen, Weifeng Ge; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16254-16263
In this paper, we develop a weakly supervised learning algorithm to learn robust semantic correspondences from large-scale datasets with only image-level labels. Following the spirit of multiple instance learning (MIL), we decompose the weakly supervised correspondence learning problem into three stages: image-level matching, region-level matching, and pixel-level matching. We propose a novel cascaded online correspondence refinement algorithm to integrate MIL and the correspondence filtering and refinement procedure into a single deep network and train this network end-to-end with only image-level supervision, i.e., without point-to-point matching information. During the correspondence learning process, pixel-to-pixel matching pairs inferred from weak supervision are propagated, filtered, and enhanced through masked correspondence voting and calibration. Besides, we design a correspondence consistency check algorithm to select images with discriminative key points to generate pseudo-labels for classical matching algorithms. Finally, we filter out about 110,000 images from the ImageNet ILSVRC training set to formulate a new dataset, called SC-ImageNet. Experiments on several popular benchmarks indicate that pre-training on SC-ImageNet can improve the performance of state-of-the-art algorithms efficiently. Our project is available on https://github.com/21210240056/SC-ImageNet.
****
Clusterformer: Cluster-based Transformer for 3D Object Detection in Point Clouds
Yu Pei, Xian Zhao, Hao Li, Jingyuan Ma, Jingwei Zhang, Shiliang Pu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6664-6673
Attributed to the unstructured and sparse nature of point clouds, the transformer shows greater potential in point clouds data processing. However, the recent query-based 3D detectors usually project the features acquired from a sparse backbone into the structured and compact Bird's Eye View(BEV) plane before adopting the transformer, which destroys the sparsity of features, introducing empty tokens and additional resource consumption for the transformer. To this end, in this paper, we propose a novel query-based 3D detector called Clusterformer, our Clusterformer regards each object as a cluster of 3D space which mainly consists of the non-empty voxels belonging to the same object, and leverages the cluster to conduct the transformer decoder to generate the proposals from the sparse voxel features directly. Such cluster-based transformer structure can effectively improve the performance and convergence speed of query-based detectors by making use of the object prior information contained in the clusters. Additionally, we in

troduce a Query2Key strategy to enhance the key and value features with the obje
ct-level information iteratively in our cluster-based transformer structure. Exp
erimental results show that the proposed Clusterformer outperforms the previous
query-based detectors with a lower latency and memory usage, which achieves stat
e-of-the-art performance on the Waymo Open Datasets and KITTI Datasets.
****
HMD-NeMo: Online 3D Avatar Motion Generation From Sparse Observations
Sadegh Aliakbarian, Fatemeh Saleh, David Collier, Pashmina Cameron, Darren Coske
r; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 9622-9631
Generating both plausible and accurate full body avatar motion is the key to the
 quality of immersive experiences in mixed reality scenarios. Head-Mounted Devic
es (HMDs) typically only provide a few input signals, such as head and hands 6-D
oF. Recently, different approaches achieved impressive performance in generating
 full body motion given only head and hands signal. However, to the best of our
knowledge, all existing approaches rely on full hand visibility. While this is t
he case when, e.g., using motion controllers, a considerable proportion of mixed
 reality experiences do not involve motion controllers and instead rely on egoce
ntric hand tracking. This introduces the challenge of partial hand visibility ow
ing to the restricted field of view of the HMD. In this paper, we propose the fi
rst unified approach, HMD-NeMo, that addresses plausible and accurate full body
motion generation even when the hands may be only partially visible. HMD-NeMo is
 a lightweight neural network that predicts the full body motion in an online an
d real-time fashion. At the heart of HMD-NeMo is the spatio-temporal encoder wit
h novel temporally adaptable mask tokens that encourage plausible motion in the
absence of hand observations. We perform extensive analysis of the impact of dif
ferent components in HMD-NeMo and introduce a new state-of-the-art on AMASS data
set through our evaluation.
****
NaviNeRF: NeRF-based 3D Representation Disentanglement by Latent Semantic Naviga
tion
Baao Xie, Bohan Li, Zequn Zhang, Junting Dong, Xin Jin, Jingyu Yang, Wenjun Zeng
; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)
, 2023, pp. 17992-18002
3D representation disentanglement aims to identify, decompose, and manipulate th
e underlying explanatory factors of 3D data, which helps AI fundamentally unders
tand our 3D world. This task is currently under-explored and poses great challen
ges: (i) the 3D representations are complex and in general contains much more in
formation than 2D image; (ii) many 3D representations are not well suited for gr
adient-based optimization, let alone disentanglement. To address these challenge
s, we use NeRF as a differentiable 3D representation, and introduce a self-super
vised Navigation to identify interpretable semantic directions in the latent spa
ce. To our best knowledge, this novel method, dubbed NaviNeRF, is the first work
 to achieve fine-grained 3D disentanglement without any priors or supervision. S
pecifically, NaviNeRF is built upon the generative NeRF pipeline, and equipped w
ith an Outer Navigation Branch and an Inner Refinement Branch. They are compleme
ntary ---- the outer navigation is to identify global-view semantic directions,
and the inner refinement dedicates to fine-grained attributes. A synergistic los
s is further devised to coordinate two branches. Extensive experiments demonstra
te that NaviNeRF has a superior fine-grained 3D disentanglement ability than the
 previous 3D-aware models. Its performance is also comparable to editing-oriente
d models relying on semantic or geometry priors.
****
Adaptive Illumination Mapping for Shadow Detection in Raw Images
Jiayu Sun, Ke Xu, Youwei Pang, Lihe Zhang, Huchuan Lu, Gerhard Hancke, Rynson La
u; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 12709-12718
Shadow detection methods rely on multi-scale contrast, especially global contras
t, information to locate shadows correctly. However, we observe that the camera
image signal processor (ISP) tends to preserve more local contrast information b

y sacrificing global contrast information during the raw-to-sRGB conversion process. This often causes existing methods to fail in scenes with high global contrast but low local contrast in shadow regions. In this paper, we propose a novel method to detect shadows from raw images. Our key idea is that instead of performing a many-to-one mapping like the ISP process, we can learn a many-to-many mapping from the high dynamic range raw images to the sRGB images of different illumination, which is able to preserve multi-scale contrast for accurate shadow detection. To this end, we first construct a new shadow dataset with 7000 raw images and shadow masks. We then propose a novel network, which includes a novel adaptive illumination mapping (AIM) module to project the input raw images into sRGB images of different intensity ranges and a shadow detection module to leverage the preserved multi-scale contrast information to detect shadows. To learn the shadow-aware adaptive illumination mapping process, we propose a novel feedback mechanism to guide the AIM during training. Experiments show that our method outperforms state-of-the-art shadow detectors. Code and dataset are available at https://github.com/jiayusun/SARA.
****
CDUL: CLIP-Driven Unsupervised Learning for Multi-Label Image Classification
Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, Song Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1348-1357
This paper presents a CLIP-based unsupervised learning method for annotation-free multi-label image classification, including three stages: initialization, training, and inference. At the initialization stage, we take full advantage of the powerful CLIP model and propose a novel approach to extend CLIP for multi-label predictions based on global-local image-text similarity aggregation. To be more specific, we split each image into snippets and leverage CLIP to generate the similarity vector for the whole image (global) as well as each snippet (local). Then a similarity aggregator is introduced to leverage the global and local similarity vectors. Using the aggregated similarity scores as the initial pseudo labels at the training stage, we propose an optimization framework to train the parameters of the classification network and refine pseudo labels for unobserved labels. During inference, only the classification network is used to predict the labels of the input image. Extensive experiments show that our method outperforms state-of-the-art unsupervised methods on MS-COCO, PASCAL VOC 2007, PASCAL VOC 2012, and NUS datasets and even achieves comparable results to weakly supervised classification methods.
****
Your Diffusion Model is Secretly a Zero-Shot Classifier
Alexander C. Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, Deepak Pathak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2206-2217
The recent wave of large-scale text-to-image diffusion models has dramatically increased our text-based image generation abilities. These models can generate realistic images for a staggering variety of prompts and exhibit impressive compositional generalization abilities. Almost all use cases thus far have solely focused on sampling; however, diffusion models can also provide conditional density estimates, which are useful for tasks beyond image generation. In this paper, we show that the density estimates from large-scale text-to-image diffusion models like Stable Diffusion can be leveraged to perform zero-shot classification without any additional training. Our generative approach to classification, which we call Diffusion Classifier, attains strong results on a variety of benchmarks and outperforms alternative methods of extracting knowledge from diffusion models. Although a gap remains between generative and discriminative approaches on zero-shot recognition tasks, our diffusion-based approach has stronger multimodal compositional reasoning abilities than competing discriminative approaches. Finally, we use Diffusion Classifier to extract standard classifiers from class-conditional diffusion models trained on ImageNet. These models approach the performance of SOTA discriminative classifiers and exhibit strong "effective robustness" to distribution shift. Overall, our results are a step toward using generative ov

er discriminative models for downstream tasks.
****
Backpropagation Path Search On Adversarial Transferability
Zhuoer Xu, Zhangxuan Gu, Jianping Zhang, Shiwen Cui, Changhua Meng, Weiqiang Wan
g; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 4663-4673
Deep neural networks are vulnerable to adversarial examples, dictating the imper
ativeness to test the model's robustness before deployment. Transfer-based attac
kers craft adversarial examples against surrogate models and transfer them to vi
ctim models deployed in the black-box situation. To enhance the adversarial tran
sferability, structure-based attackers adjust the backpropagation path to avoid
the attack from overfitting the surrogate model. However, existing structure-bas
ed attackers fail to explore the convolution module in CNNs and modify the backp
ropagation graph heuristically, leading to limited effectiveness. In this paper,
 we propose backPropagation pAth Search (PAS), solving the aforementioned two pr
oblems. We first propose SkipConv to adjust the backpropagation path of convolut
ion by structural reparameterization. To overcome the drawback of heuristically
designed backpropagation paths, we further construct a DAG-based search space, u
tilize one-step approximation for path evaluation and employ Bayesian Optimizati
on to search for the optimal path. We conduct comprehensive experiments in a wid
e range of transfer settings, showing that PAS improves the attack success rate
by a huge margin for both normally trained and defense models.
****
Boosting Adversarial Transferability via Gradient Relevance Attack
Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, Wuming Jiang; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4741-
4750
Plentiful adversarial attack researches have revealed the fragility of deep neur
al networks (DNNs), where the imperceptible perturbations can cause drastic chan
ges in the output. Among the diverse types of attack methods, gradient-based att
acks are powerful and easy to implement, arousing wide concern for the security
problem of DNNs. However, under the black-box setting, the existing gradient-bas
ed attacks have much trouble in breaking through DNN models with defense technol
ogies, especially those adversarially trained models. To make adversarial exampl
es more transferable, in this paper, we explore the fluctuation phenomenon on th
e plus-minus sign of the adversarial perturbations' pixels during the generation
 of adversarial examples, and propose an ingenious Gradient Relevance Attack (GR
A). Specifically, two gradient relevance frameworks are presented to better util
ize the information in the neighborhood of the input, which can correct the upda
te direction adaptively. Then we adjust the update step at each iteration with a
 decay indicator to counter the fluctuation. Experiment results on a subset of t
he ILSVRC 2012 validation set forcefully verify the effectiveness of GRA. Furthe
rmore, the attack success rates of 68.7% and 64.8% on Tencent Cloud and Baidu AI
 Cloud further indicate that GRA can craft adversarial examples with the ability
 to transfer across both datasets and model architectures. Code is released at h
ttps://github.com/RYC-98/GRA.
****
Image-Free Classifier Injection for Zero-Shot Classification
Anders Christensen, Massimiliano Mancini, A. Sophia Koepke, Ole Winther, Zeynep
Akata; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2023, pp. 19072-19081
Zero-shot learning models achieve remarkable results on image classification for
 samples from classes that were not seen during training. However, such models m
ust be trained from scratch with specialised methods: therefore, access to a tra
ining dataset is required when the need for zero-shot classification arises. In
this paper, we aim to equip pre-trained models with zero-shot classification cap
abilities without the use of image data. We achieve this with our proposed Image
-free Classifier Injection with Semantics (ICIS) that injects classifiers for ne
w, unseen classes into pre-trained classification models in a post-hoc fashion w
ithout relying on image data. Instead, the existing classifier weights and simpl

e class-wise descriptors, such as class names or attributes, are used. ICIS has two encoder-decoder networks that learn to reconstruct classifier weights from descriptors (and vice versa), exploiting (cross-)reconstruction and cosine losses to regularise the decoding process. Notably, ICIS can be cheaply trained and applied directly on top of pre-trained classification models. Experiments on benchmark ZSL datasets show that ICIS produces unseen classifier weights that achieve strong (generalised) zero-shot classification performance. Code is available at https://github.com/ExplainableML/ImageFreeZSL.
****
CLIPN for Zero-Shot OOD Detection: Teaching CLIP to Say No
Hualiang Wang, Yi Li, Huifeng Yao, Xiaomeng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1802-1812
Out-of-distribution (OOD) detection refers to training the model on in-distribution (ID) dataset to classify if the input images come from unknown classes. Considerable efforts have been invested in designing various OOD detection methods based on either convolutional neural networks or transformers. However, Zero-shot OOD detection methods driven by CLIP, which require only class names for ID, have received less attention. This paper presents a novel method, namely CLIP saying no (CLIPN), which empowers "no" logic within CLIP. Our key motivation is to equip CLIP with the capability of distinguishing OOD and ID samples via positive-semantic prompts and negation-semantic prompts. To be specific, we design a novel learnable "no" prompt and a "no" text encoder to capture the negation-semantic with images. Subsequently, we introduce two loss functions: the image-text binary-opposite loss and the text semantic-opposite loss, which we use to teach CLIPN to associate images with "no" prompts, thereby enabling it to identify unknown samples. Furthermore, we propose two threshold-free inference algorithms to perform OOD detection via using negation semantics from "no" prompts and text encoder. Experimental results on 9 benchmark datasets (3 ID datasets and 6 OOD datasets) for the OOD detection task demonstrate that CLIPN outperforms 7 well-used algorithms by at least 1.1% and 7.37% on AUROC and FPR95 on zero-shot OOD detection of ImageNet-1K. Our CLIPN can serve as a solid foundation for leveraging CLIP effectively in downstream OOD tasks.
****
CO-Net: Learning Multiple Point Cloud Tasks at Once with A Cohesive Network
Tao Xie, Ke Wang, Siyi Lu, Yukun Zhang, Kun Dai, Xiaoyu Li, Jie Xu, Li Wang, Lijun Zhao, Xinyu Zhang, Ruifeng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3523-3533
We present CO-Net, a cohesive framework that optimizes multiple point cloud tasks collectively across heterogeneous dataset domains. CO-Net maintains the characteristics of high storage efficiency since models with the preponderance of shared parameters can be assembled into a single model. Specifically, we leverage residual MLP (Res-MLP) block for effective feature extraction and scale it gracefully along the depth and width of the network to meet the demands of different tasks. Based on the block, we propose a novel nested layer-wise processing policy, which identifies the optimal architecture for each task while provides partial sharing parameters and partial non-sharing parameters inside each layer of the block. Such policy tackles the inherent challenges of multi-task learning on point cloud, e.g., diverse model topologies resulting from task skew and conflicting gradients induced by heterogeneous dataset domains. Finally, we propose a sign-based gradient surgery to promote the training of CO-Net, thereby emphasizing the usage of task-shared parameters and guaranteeing that each task can be thoroughly optimized. Experimental results reveal that models optimized by CO-Net jointly for all point cloud tasks maintain much fewer computation cost and overall storage cost yet outpace prior methods by a significant margin. We also demonstrate that CO-Net allows incremental learning and prevents catastrophic amnesia when adapting to a new point cloud task.
****
Quality Diversity for Visual Pre-Training
Ruchika Chavhan, Henry Gouk, Da Li, Timothy Hospedales; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5384-5394

Models pre-trained on large datasets such as ImageNet provide the de-facto standard for transfer learning, with both supervised and self-supervised approaches proving effective. However, emerging evidence suggests that any single pre-trained feature will not perform well on diverse downstream tasks. Each pre-training strategy encodes a certain inductive bias, which may suit some downstream tasks but not others. Notably, the augmentations used in both supervised and self-supervised training lead to features with high invariance to spatial and appearance transformations. This renders them sub-optimal for tasks that demand sensitivity to these factors. In this paper we develop a feature that better supports diverse downstream tasks by providing a diverse set of sensitivities and invariances. In particular, we are inspired by Quality-Diversity in evolution, to define a pre-training objective that requires high quality yet diverse features -- where diversity is defined in terms of transformation (in)variances. Our framework plugs in to both supervised and self-supervised pre-training, and produces a small ensemble of features. We further show how downstream tasks can easily and efficiently select their preferred (in)variances. Both empirical and theoretical analysis show the efficacy of our representation and transfer learning approach for diverse downstream tasks.
****
UniDexGrasp++: Improving Dexterous Grasping Policy Learning via Geometry-Aware Curriculum and Iterative Generalist-Specialist Learning
Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, He Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3891-3902
We propose a novel, object-agnostic method for learning a universal policy for dexterous object grasping from realistic point cloud observations and proprioceptive information under a table-top setting, namely UniDexGrasp++. To address the challenge of learning the vision-based policy across thousands of object instances, we propose Geometry-aware Curriculum Learning (GeoCurriculum) and Geometry-aware iterative Generalist-Specialist Learning (GiGSL) which leverage the geometry feature of the task and significantly improve the generalizability. With our proposed techniques, our final policy shows universal dexterous grasping on thousands of object instances with 85.4% and 78.2% success rate on the train set and test set which outperforms the state-of-the-art baseline UniDexGrasp by 11.7% and 11.3%, respectively.
****
Multi-Scale Residual Low-Pass Filter Network for Image Deblurring
Jiangxin Dong, Jinshan Pan, Zhongbao Yang, Jinhui Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12345-12354
We present a simple and effective Multi-scale Residual Low-Pass Filter Network (MRLPFNet) that jointly explores the image details and main structures for image deblurring. Our work is motivated by an observation that the difference between the blurry image and the clear one not only contains high-frequency contents(Note that the high-frequency contents in an image correspond to the image details, while the low-frequency ones denote the main structures of an image.) but also includes low-frequency information due to the influence of blur, while using the standard residual learning is less effective for modeling the main structure distorted by the blur. Considering that the low-frequency contents usually correspond to main global structures that are spatially variant, we first propose a learnable low-pass filter based on a self-attention mechanism to adaptively explore the global contexts for better modeling the low-frequency information. Then we embed it into a Residual Low-Pass Filter (RLPF) module, which involves an additional fully convolutional neural network with the standard residual learning to model the high-frequency information. We formulate the RLPF module into an end-to-end trainable network based on an encoder and decoder architecture and develop a wavelet-based feature fusion to fuse the multi-scale features. Experimental results show that our method performs favorably against state-of-the-art ones on commonly-used benchmarks.
****
FerKD: Surgical Label Adaptation for Efficient Distillation

Zhiqiang Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1666-1675
We present FerKD, a novel efficient knowledge distillation framework that incorporates partial soft-hard label adaptation coupled with a region-calibration mechanism. Our approach stems from the observation and intuition that standard data augmentations, such as RandomResizedCrop, tend to transform inputs into diverse conditions: easy positives, hard positives, or hard negatives. In traditional distillation frameworks, these transformed samples are utilized equally through their predictive probabilities derived from pretrained teacher models. However, merely relying on prediction values from a pretrained teacher, a common practice in prior studies, neglects the reliability of these soft label predictions. To address this, we propose a new scheme that calibrates the less-confident regions to be the context using softened hard groundtruth labels. Our approach involves the processes of hard regions mining + calibration. We demonstrate empirically that this method can dramatically improve the convergence speed and final accuracy. Additionally, we find that a consistent mixing strategy can stabilize the distributions of soft supervision, taking advantage of the soft labels. As a result, we introduce a stabilized SelfMix augmentation that weakens the variation of the mixed images and corresponding soft labels through mixing similar regions within the same image. FerKD is an intuitive and well-designed learning system that eliminates several heuristics and hyperparameters in former FKD solution. More importantly, it achieves remarkable improvement on ImageNet-1K and downstream tasks. For instance, FerKD achieves 81.2% on ImageNet-1K with ResNet-50, outperforming FKD and FunMatch by remarkable margins. Leveraging better pre-trained weights and larger architectures, our finetuned ViT-G14 even achieves 89.9%. Our code is available at https://github.com/szq0214/FKD/tree/main/FerKD.
****
Neural Fields for Structured Lighting
Aarrushi Shandilya, Benjamin Attal, Christian Richardt, James Tompkin, Matthew O'toole; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3512-3522
We present an image formation model and optimization procedure that combines the advantages of neural radiance fields and structured light imaging. Existing depth-supervised neural models rely on depth sensors to accurately capture the scene's geometry. However, the depth maps recovered by these sensors can be prone to error, or even fail outright. Instead of depending on the fidelity of processed depth maps from a structured light system, a more principled approach is to explicitly model the raw structured light images themselves. Our proposed approach enables the estimation of high-fidelity depth maps, including for objects with complex material properties (e.g., partially-transparent surfaces). Besides computing depth, the raw structured light images also confer other useful radiometric cues, which enable predicting surface normals and decomposing scene appearance in terms of a direct, indirect, and ambient component. We evaluate our framework quantitatively and qualitatively on a range of real and synthetic scenes, and decompose scenes into their constituent components for novel views.
****
ClothPose: A Real-world Benchmark for Visual Analysis of Garment Pose via An Indirect Recording Solution
Wenqiang Xu, Wenxin Du, Han Xue, Yutong Li, Ruolin Ye, Yan-Feng Wang, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 58-68
Garments are important and pervasive in daily life. However, visual analysis on them for pose estimation is challenging because it requires recovering the complete configurations of garments, which is difficult, if not impossible, to annotate in the real world. In this work, we propose a recording system, GarmentTwin, which can track garment poses in dynamic settings such as manipulation. GarmentTwin first collects garment models and RGB-D manipulation videos from the real world and then replays the manipulation process using physics-based animation. This way, we can obtain deformed garments with poses coarsely aligned with real-world observations. Finally, we adopt an optimization-based approach to fit the pos

e with real-world observations. We verify the fitting results quantitatively and qualitatively. With GarmentTwin, we construct a large-scale dataset named ClothPose, which consists of 30K RGB-D frames from 2K video clips on 600 garments of 10 categories. We benchmark two tasks on the proposed ClothPose: non-rigid reconstruction and pose estimation. The experiments show that previous baseline methods struggle with highly large non-rigid deformation of manipulated garments. Therefore, we hope that the recording system and the dataset can facilitate research on pose estimation tasks on non-rigid objects. Datasets, models, and codes are made publicly available.
****
Semantically Structured Image Compression via Irregular Group-Based Decoupling
Ruoyu Feng, Yixin Gao, Xin Jin, Runsen Feng, Zhibo Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17237-17247
Image compression techniques typically focus on compressing rectangular images for human consumption, however, resulting in transmitting redundant content for downstream applications. To overcome this limitation, some previous works propose to semantically structure the bitstream, which can meet specific application requirements by selective transmission and reconstruction. Nevertheless, they divide the input image into multiple rectangular regions according to semantics and ignore avoiding information interaction among them, causing waste of bitrate and distorted reconstruction of region boundaries. In this paper, we propose to decouple an image into multiple groups with irregular shapes based on a customized group mask and compress them independently. Our group mask describes the image at a finer granularity, enabling significant bitrate saving by reducing the transmission of redundant content. Moreover, to ensure the fidelity of selective reconstruction, this paper proposes the concept of group-independent transform that maintain the independence among distinct groups. And we instantiate it by the proposed Group-Independent Swin-Block (GI Swin-Block). Experimental results demonstrate that our framework structures the bitstream with negligible cost, and exhibits superior performance on both visual quality and intelligent task supporting.
****
PhaseMP: Robust 3D Pose Estimation via Phase-conditioned Human Motion Prior
Mingyi Shi, Sebastian Starke, Yuting Ye, Taku Komura, Jungdam Won; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14725-14737
We present a novel motion prior, called PhaseMP, modeling a probability distribution on pose transitions conditioned by a frequency domain feature extracted from a periodic autoencoder. The phase feature further enforces the pose transitions to be unidirectional (i.e. no backward movement in time), from which more stable and natural motions can be generated. Specifically, our motion prior can be useful for accurately estimating 3D human motions in the presence of challenging input data, including long periods of spatial and temporal occlusion, as well as noisy sensor measurements. Through a comprehensive evaluation, we demonstrate the efficacy of our novel motion prior, showcasing its superiority over existing state-of-the-art methods by a significant margin across various applications, including video-to-motion and motion estimation from sparse sensor data, and etc.
****
NLOS-NeuS: Non-line-of-sight Neural Implicit Surface
Yuki Fujimura, Takahiro Kushida, Takuya Funatomi, Yasuhiro Mukaigawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10532-10541
Non-line-of-sight (NLOS) imaging is conducted to infer invisible scenes from indirect light on visible objects. The neural transient field (NeTF) was proposed for representing scenes as neural radiance fields in NLOS scenes. We propose NLOS neural implicit surface (NLOS-NeuS), which extends the NeTF to neural implicit surfaces with a signed distance function (SDF) for reconstructing three-dimensional surfaces in NLOS scenes. We introduce two constraints as loss functions for correctly learning an SDF to avoid non-zero level-set surfaces. We also introduce a lower bound constraint of an SDF based on the geometry of the first-returnin

g photons. The experimental results indicate that these constraints are essential for learning a correct SDF in NLOS scenes. Compared with previous methods with discretized representation, NLOS-NeuS with the neural continuous representation enables us to reconstruct smooth surfaces while preserving fine details in NLOS scenes. To the best of our knowledge, this is the first study on neural implicit surfaces with volume rendering in NLOS scenes. Project page: https://yfujimura. github.io/nlos-neus/
****
Unsupervised Object Localization with Representer Point Selection
Yeonghwan Song, Seokwoo Jang, Dina Katabi, Jeany Son; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6534-6544
We propose a novel unsupervised object localization method that allows us to explain the predictions of the model by utilizing self-supervised pre-trained models without additional finetuning. Existing unsupervised and self-supervised object localization methods often utilize class-agnostic activation maps or self-similarity maps of a pre-trained model. Although these maps can offer valuable information for localization, their limited ability to explain how the model makes predictions remains challenging. In this paper, we propose a simple yet effective unsupervised object localization method based on representer point selection, where the predictions of the model can be represented as a linear combination of representer values of training points. By selecting representer points, which are the most important examples for the model predictions, our model can provide insights into how the model predicts the foreground object by providing relevant examples as well as their importance. Our method outperforms the state-of-the-art unsupervised and self-supervised object localization methods on various datasets with significant margins and even outperforms recent weakly supervised and few-shot methods.
****
SEMPART: Self-supervised Multi-resolution Partitioning of Image Semantics
Sriram Ravindran, Debraj Basu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 723-733
Accurately determining salient regions of an image is challenging when labeled data is scarce. DINO-based self-supervised approaches have recently leveraged meaningful image semantics captured by patch-wise features for locating foreground objects. Recent methods have also incorporated intuitive priors and demonstrated value in unsupervised methods for object partitioning. In this paper, we propose SEMPART, which jointly infers coarse and fine bi-partitions over an image's DINO-based semantic graph. Furthermore, SEMPART preserves fine boundary details using graph-driven regularization and successfully distills the coarse mask semantics into the fine mask. Our salient object detection and single object localization findings suggest that SEMPART produces high-quality masks rapidly without additional post-processing and benefits from co-optimizing the coarse and fine branches.
****
Flatness-Aware Minimization for Domain Generalization
Xingxuan Zhang, Renzhe Xu, Han Yu, Yancheng Dong, Pengfei Tian, Peng Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5189-5202
Domain generalization (DG) seeks to learn robust models that generalize well under unknown distribution shifts. As a critical aspect of DG, optimizer selection has not been explored in depth. Currently, most DG methods follow the widely used benchmark, DomainBed, and utilize Adam as the default optimizer for all datasets. However, we reveal that Adam is not necessarily the optimal choice for the majority of current DG methods and datasets. Based on the perspective of loss landscape flatness, we propose a novel approach, Flatness-Aware Minimization for Domain Generalization (FAD), which can efficiently optimize both zeroth-order and first-order flatness simultaneously for DG. We provide theoretical analyses of the FAD's out-of-distribution (OOD) generalization error and convergence. Our experimental results demonstrate the superiority of FAD on various DG datasets.
****

ProtoFL: Unsupervised Federated Learning via Prototypical Distillation
Hansol Kim, Youngjun Kwak, Minyoung Jung, Jinho Shin, Youngsung Kim, Changick Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6470-6479
Federated learning (FL) is a promising approach for enhancing data privacy preservation, particularly for authentication systems. However, limited round communications, scarce representation, and scalability pose significant challenges to its deployment, hindering its full potential. In this paper, we propose 'ProtoFL', Prototypical Representation Distillation based unsupervised Federated Learning to enhance the representation power of a global model and reduce round communication costs. Additionally, we introduce a local one-class classifier based on normalizing flows to improve performance with limited data. Our study represents the first investigation of using FL to improve one-class classification performance. We conduct extensive experiments on five widely used benchmarks, namely MNIST, CIFAR-10, CIFAR-100, ImageNet-30, and Keystroke-Dynamics, to demonstrate the superior performance of our proposed framework over previous methods in the literature.
****
Augmenting and Aligning Snippets for Few-Shot Video Domain Adaptation
Yuecong Xu, Jianfei Yang, Yunjiao Zhou, Zhenghua Chen, Min Wu, Xiaoli Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13445-13456
For video models to be transferred and applied seamlessly across video tasks in varied environments, Video Unsupervised Domain Adaptation (VUDA) has been introduced to improve the robustness and transferability of video models. However, current VUDA methods rely on a vast amount of high-quality unlabeled target data, which may not be available in real-world cases. We thus consider a more realistic Few-Shot Video-based Domain Adaptation (FSVDA) scenario where we adapt video models with only a few target video samples. While a few methods have touched upon Few-Shot Domain Adaptation (FSDA) in images and in FSVDA, they rely primarily on spatial augmentation for target domain expansion with alignment performed statistically at the instance level. However, videos contain more knowledge in terms of rich temporal and semantic information, which should be fully considered while augmenting target domains and performing alignment in FSVDA. We propose a novel SSA2lign to address FSVDA at the snippet level, where the target domain is expanded through a simple snippet-level augmentation followed by the attentive alignment of snippets both semantically and statistically, where semantic alignment of snippets is conducted through multiple perspectives. Empirical results demonstrate state-of-the-art performance of SSA2lign across multiple cross-domain action recognition benchmarks.
****
Self-Organizing Pathway Expansion for Non-Exemplar Class-Incremental Learning
Kai Zhu, Kecheng Zheng, Ruili Feng, Deli Zhao, Yang Cao, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19204-19213
Non-exemplar class-incremental learning aims to recognize both the old and new classes without access to old class samples. The conflict between old and new class optimization is exacerbated since the shared neural pathways can only be differentiated by the incremental samples. To address this problem, we propose a novel self-organizing pathway expansion scheme. Our scheme consists of a class-specific pathway organization strategy that reduces the coupling of optimization pathway among different classes to enhance the independence of the feature representation, and a pathway-guided feature optimization mechanism to mitigate the update interference between the old and new classes. Extensive experiments on four datasets demonstrate significant performance gains, outperforming the state-of-the-art methods by a margin of 1%, 3%, 2% and 2%, respectively.
****
Preserving Tumor Volumes for Unsupervised Medical Image Registration
Qihua Dong, Hao Du, Ying Song, Yan Xu, Jing Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21208-21218

Medical image registration is a critical task that estimates the spatial correspondence between pairs of images. However, current traditional and learning-based methods rely on similarity measures to generate a deforming field, which often results in disproportionate volume changes in dissimilar regions, especially in tumor regions. These changes can significantly alter the tumor size and underlying anatomy, which limits the practical use of image registration in clinical diagnosis. To address this issue, we have formulated image registration with tumors as a constraint problem that preserves tumor volumes while maximizing image similarity in other normal regions. Our proposed framework involves a two-stage process. In the first stage, we use similarity-based registration to identify potential tumor regions by their volume change, generating a soft tumor mask accordingly. In the second stage, we propose a volume-preserving registration with a novel adaptive volume-preserving loss that penalizes the change in size adaptively based on the masks calculated from the previous stage. Our approach balances image similarity and volume preservation in different regions, i.e., normal and tumor regions, by using soft tumor masks to adjust the imposition of volume-preserving loss on each one. This ensures that the tumor volume is preserved during the registration process. We have evaluated our framework on various datasets and network architectures, demonstrating that our method successfully preserves the tumor volume while achieving comparable registration results with state-of-the-art methods. Our code is at: https://dddraxxx.github.io/Volume-Preserving-Registration/.
****
Multi-label Affordance Mapping from Egocentric Vision
Lorenzo Mur-Labadia, Jose J. Guerrero, Ruben Martinez-Cantin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5238-5249
Accurate affordance detection and segmentation with pixel precision is an important piece in many complex systems based on interactions, such as robots and assitive devices. We present a new approach to affordance perception which enables accurate multi-label segmentation. Our approach can be used to automatically annotate grounded affordances from first person videos of interactions using a 3D map of the environment providing pixel level precision for the affordance location. We use this method to build the largest and most complete dataset on affordances based on the EPIC-Kitchen dataset, EPIC-Aff, which provides automatic, interaction-grounded, multi-label, metric and spatial affordance annotations. Then, we propose a new approach to affordance segmentation based on multi-label detection which enables multiple affordances to co-exists in the same space, for example if they are associated with the same object. We present several strategies of multi-label detection using several segmentation architectures. The experimental results highlights the importance of the multi-label detection. Finally, we show how our metric representation can be exploited for build a map of interaction hotspots in spatial action-centric zones and use that representation to perform a task-oriented navigation.
****
Towards Real-World Burst Image Super-Resolution: Benchmark and Method
Pengxu Wei, Yujing Sun, Xingbei Guo, Chang Liu, Guanbin Li, Jie Chen, Xiangyang Ji, Liang Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13233-13242
Despite substantial advances, single-image super-resolution (SISR) is always in a dilemma to reconstruct high-quality images with limited information from one input image, especially in realistic scenarios. In this paper, we establish a large-scale real-world burst super-resolution dataset, i.e., RealBSR, to explore the faithful reconstruction of image details from multiple frames. Furthermore, we introduce a Federated Burst Affinity network (FBAnet) to investigate non-trivial pixel-wise displacements among images under real-world image degradation. Specifically, rather than using pixel-wise alignment, our FBAnet employs a simple homography alignment from a structural geometry aspect and a Federated Affinity Fusion (FAF) strategy to aggregate the complementary information among frames. Those fused informative representations are fed to a Transformer-based module of bu

rst representation decoding. Besides, we have conducted extensive experiments on two versions of our datasets, i.e., RealBSR-RAW and RealBSR-RGB. Experimental results demonstrate that our FBAnet outperforms existing state-of-the-art burst SR methods and also achieves visually-pleasant SR image predictions with model details. Our dataset, codes, and models are publicly available at https://github.com/yjsunnn/FBANet.
****
Unified Adversarial Patch for Cross-Modal Attacks in the Physical World
Xingxing Wei, Yao Huang, Yitong Sun, Jie Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4445-4454
Recently, physical adversarial attacks have been presented to evade DNNs-based object detectors. To ensure the security, many scenarios are simultaneously deployed with visible sensors and infrared sensors, leading to the failures of these single-modal physical attacks. To show the potential risks under such scenes, we propose a unified adversarial patch to perform cross-modal physical attacks, i.e., fooling visible and infrared object detectors at the same time via a single patch. Considering different imaging mechanisms of visible and infrared sensors, our work focuses on modeling the shapes of adversarial patches, which can be captured in different modalities when they change. To this end, we design a novel boundary-limited shape optimization to achieve the compact and smooth shapes, and thus they can be easily implemented in the physical world. In addition, to balance the fooling degree between visible detector and infrared detector during the optimization process, we propose a score-aware iterative evaluation, which can guide the adversarial patch to iteratively reduce the predicted scores of the multi-modal sensors. We finally test our method against the one-stage detector: YOLOv3 and the two-stage detector: Faster RCNN. Results show that our unified patch achieves an Attack Success Rate (ASR) of 73.33% and 69.17%, respectively. More importantly, we verify the effective attacks in the physical world when visible and infrared sensors shoot the objects under various settings like different angles, distances, postures, and scenes.
****
Unsupervised Accuracy Estimation of Deep Visual Models using Domain-Adaptive Adversarial Perturbation without Source Samples
JoonHo Lee, Jae Oh Woo, Hankyu Moon, Kwonho Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16443-16452
Deploying deep visual models can lead to performance drops due to the discrepancies between source and target distributions. Several approaches leverage labeled source data to estimate target domain accuracy, but accessing labeled source data is often prohibitively difficult due to data confidentiality or resource limitations on serving devices. Our work proposes a new framework to estimate model accuracy on unlabeled target data without access to source data. We investigate the feasibility of using pseudo-labels for accuracy estimation and evolve this idea into adopting recent advances in source-free domain adaptation algorithms. Our approach measures the disagreement rate between the source hypothesis and the target pseudo-labeling function, adapted from the source hypothesis. We mitigate the impact of erroneous pseudo-labels that may arise due to a high ideal joint hypothesis risk by employing adaptive adversarial perturbation on the input of the target model. Our proposed source-free framework effectively addresses the challenging distribution shift scenarios and outperforms existing methods requiring source data and labels for training.
****
Misalign, Contrast then Distill: Rethinking Misalignments in Language-Image Pre-training
Bumsoo Kim, Yeonsik Jo, Jinhyung Kim, Seunghwan Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2563-2572
Contrastive Language-Image Pretraining has emerged as a prominent approach for training vision and text encoders with uncurated image-text pairs from the web. To enhance data-efficiency, recent efforts have introduced additional supervision terms that involve random-augmented views of the image. However, since the image augmentation process is unaware of its text counterpart, this procedure could

cause various degrees of image-text misalignments during training. Prior methods either disregarded this discrepancy or introduced external models to mitigate the impact of misalignments during training. In contrast, we propose a novel metric learning approach that capitalizes on these misalignments as an additional training source, which we term "Misalign, Contrast then Distill (MCD)". Unlike previous methods that treat augmented images and their text counterparts as simple positive pairs, MCD predicts the continuous scales of misalignment caused by the augmentation. Our extensive experimental results show that our proposed MCD achieves state-of-the-art transferability in multiple classification and retrieval downstream datasets.
****
SYENet: A Simple Yet Effective Network for Multiple Low-Level Vision Tasks with Real-Time Performance on Mobile Device
Weiran Gou, Ziyao Yi, Yan Xiang, Shaoqing Li, Zibin Liu, Dehui Kong, Ke Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12182-12195
With the rapid development of AI hardware accelerators, applying deep learning-based algorithms to solve various low-level vision tasks on mobile devices has gradually become possible. However, two main problems still need to be solved. Firstly, most low-level vision algorithms are task-specific and independent to each other, which makes them difficult to integrate into a single neural network architecture and accelerate simultaneously without task-level time-multiplexing. Secondly, most of these networks feature large amounts of parameters and huge computational costs in terms of multiplication-and-accumulation operations, and thus it is difficult to achieve real-time performance, especially on mobile devices with limited computing power. To tackle with these problems, we propose a novel network, SYENet, with only 6K parameters. The SYENet consists of two asymmetrical branches with simple building blocks and is able to handle multiple low-level vision tasks on mobile devices in a real-time manner. To effectively connect the results by asymmetrical branches, a Quadratic Connection Unit(QCU) is proposed. Furthermore, in order to improve visual quality, a new Regression Focal Loss is proposed to process the image. The proposed method proves its
superior performance with the best PSNR and visual quality as compared with other networks in real-time applications such as Image Signal Processing(ISP), Low-Light Enhancement(LLE), and Super-Resolution(SR) with 2K60FPS throughput on Qualcomm 8 Gen 1 mobile SoC(System-on-Chip). Particularly, for ISP task, SYENet got the highest score in MAI 2022 Learned Smartphone ISP challenge.
****
MATE: Masked Autoencoders are Online 3D Test-Time Learners
M. Jehanzeb Mirza, Inkyu Shin, Wei Lin, Andreas Schriebl, Kunyang Sun, Jaesung Choe, Mateusz Kozinski, Horst Possegger, In So Kweon, Kuk-Jin Yoon, Horst Bischof; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16709-16718
Our MATE is the first Test-Time-Training (TTT) method designed for 3D data, which makes deep networks trained for point cloud classification robust to distribution shifts occurring in test data. Like existing TTT methods from the 2D image domain, MATE also leverages test data for adaptation. Its test-time objective is that of a Masked Autoencoder: a large portion of each test point cloud is removed before it is fed to the network, tasked with reconstructing the full point cloud. Once the network is updated, it is used to classify the point cloud. We test MATE on several 3D object classification datasets and show that it significantly improves robustness of deep networks to several types of corruptions commonly occurring in 3D point clouds. We show that MATE is very efficient in terms of the fraction of points it needs for the adaptation. It can effectively adapt given as few as 5% of tokens of each test sample, making it extremely lightweight. Our experiments show that MATE also achieves competitive performance by adapting sparsely on the test data, which further reduces its computational overhead, making it ideal for real-time applications.
****
EdaDet: Open-Vocabulary Object Detection Using Early Dense Alignment

Cheng Shi, Sibei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15724-15734

Vision-language models such as CLIP have boosted the performance of open-vocabulary object detection, where the detector is trained on base categories but required to detect novel categories. Existing methods leverage CLIP's strong zero-shot recognition ability to align object-level embeddings with textual embeddings of categories. However, we observe that using CLIP for object-level alignment results in overfitting to base categories, i.e., novel categories most similar to base categories have particularly poor performance as they are recognized as similar base categories. In this paper, we first identify that the loss of critical fine-grained local image semantics hinders existing methods from attaining strong base-to-novel generalization. Then, we propose Early Dense Alignment (EDA) to bridge the gap between generalizable local semantics and object-level prediction. In EDA, we use object-level supervision to learn the dense-level rather than object-level alignment to maintain the local fine-grained semantics. Extensive experiments demonstrate our superior performance to competing approaches under the same strict setting and without using external training resources, i.e., improving the +8.4% novel box AP50 on COCO and +3.9% rare mask AP on LVIS.

****

MixPath: A Unified Approach for One-shot Neural Architecture Search

Xiangxiang Chu, Shun Lu, Xudong Li, Bo Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5972-5981

Blending multiple convolutional kernels is proved advantageous in neural architecture design. However, current two-stage neural architecture search methods are mainly limited to single-path search spaces. How to efficiently search models of multi-path structures remains a difficult problem. In this paper, we are motivated to train a one-shot multi-path supernet to accurately evaluate the candidate architectures. Specifically, we discover that in the studied search spaces, feature vectors summed from multiple paths are nearly multiples of those from a single path. Such disparity perturbs the supernet training and its ranking ability. Therefore, we propose a novel mechanism called Shadow Batch Normalization (SBN) to regularize the disparate feature statistics. Extensive experiments prove that SBNs are capable of stabilizing the optimization and improving ranking performance. We call our unified multi-path one-shot approach as MixPath, which generates a series of models that achieve state-of-the-art results on ImageNet.

****

Enhancing NeRF akin to Enhancing LLMs: Generalizable NeRF Transformer with Mixture-of-View-Experts

Wenyan Cong, Hanxue Liang, Peihao Wang, Zhiwen Fan, Tianlong Chen, Mukund Varma, Yi Wang, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3193-3204

Cross-scene generalizable NeRF models, which can directly synthesize novel views of unseen scenes, have become a new spotlight of the NeRF field. Several existing attempts rely on increasingly end-to-end "neuralized" architectures, i.e., replacing scene representation and/or rendering modules with performant neural networks such as transformers, and turning novel view synthesis into a feed-forward inference pipeline. While those feedforward "neuralized" architectures still do not fit diverse scenes well out of the box, we propose to bridge them with the powerful Mixture-of-Experts (MoE) idea from large language models (LLMs), which has demonstrated superior generalization ability by balancing between larger overall model capacity and flexible per-instance specialization. Starting from a recent generalizable NeRF architecture called GNT, we first demonstrate that MoE can be neatly plugged in to enhance the model. We further customize a shared permanent expert and a geometry-aware consistency loss to enforce cross-scene consistency and spatial smoothness respectively, which are essential for generalizable view synthesis. Our proposed model, dubbed GNT with Mixture-of-View-Experts (GNT-MOVE), has experimentally shown state-of-the-art results when transferring to unseen scenes, indicating remarkably better cross-scene generalization in both zero-shot and few-shot settings. Our codes are available at https://github.com/VITA-Group/GNT-MOVE.

****
Task-aware Adaptive Learning for Cross-domain Few-shot Learning
Yurong Guo, Ruoyi Du, Yuan Dong, Timothy Hospedales, Yi-Zhe Song, Zhanyu Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1590-1599
Although existing few-shot learning works yield promising results for in-domain queries, they still suffer from weak cross-domain generalization. Limited support data requires effective knowledge transfer, but domain-shift makes this harder. Towards this emerging challenge, researchers improved adaptation by introducing task-specific parameters, which are directly optimized and estimated for each task. However, adding a fixed number of additional parameters fails to consider the diverse domain shifts between target tasks and the source domain, limiting efficacy. In this paper, we first observe the dependence of task-specific parameter configuration on the target task. Abundant task-specific parameters may over-fit, and insufficient task-specific parameters may result in under-adaptation -- but the optimal task-specific configuration varies for different test tasks. Based on these findings, we propose the Task-aware Adaptive Network (TA2-Net), which is trained by reinforcement learning to adaptively estimate the optimal task-specific parameter configuration for each test task. It learns, for example, that tasks with significant domain shift usually have a larger need for task-specific parameters for adaptation. We evaluate our model on Meta-dataset. Empirical results show that our model outperforms existing state-of-the-art methods.
****
Two Birds, One Stone: A Unified Framework for Joint Learning of Image and Video Style Transfers
Bohai Gu, Heng Fan, Libo Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23545-23554
Current arbitrary style transfer models are limited to either image or video domains. In order to achieve satisfying image and video style transfers, two different models are inevitably required with separate training processes on image and video domains, respectively. In this paper, we show that this can be precluded by introducing UniST, a Unified Style Transfer framework for both images and videos. At the core of UniST is a domain interaction transformer (DIT), which first explores context information within the specific domain and then interacts contextualized domain information for joint learning. In particular, DIT enables exploration of temporal information from videos for the image style transfer task and meanwhile allows rich appearance texture from images for video style transfer, thus leading to mutual benefits. Considering heavy computation of traditional multi-head self-attention, we present a simple yet effective axial multi-head self-attention (AMSA) for DIT, which improves computational efficiency while maintains style transfer performance. To verify the effectiveness of UniST, we conduct extensive experiments on both image and video style transfer tasks and show that UniST performs favorably against state-of-the-art approaches on both tasks. Code is available at https://github.com/NevSNev/UniST.
****
Revisiting Domain-Adaptive 3D Object Detection by Reliable, Diverse and Class-balanced Pseudo-Labeling
Zhuoxiao Chen, Yadan Luo, Zheng Wang, Mahsa Baktashmotlagh, Zi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3714-3726
Unsupervised domain adaptation (DA) with the aid of pseudo labeling techniques has emerged as a crucial approach for domain-adaptive 3D object detection. While effective, existing DA methods suffer from a substantial drop in performance when applied to a multi-class training setting, due to the co-existence of low-quality pseudo labels and class imbalance issues. In this paper, we address this challenge by proposing a novel ReDB framework tailored for learning to detect all classes at once. Our approach produces Reliable, Diverse, and class-Balanced pseudo 3D boxes to iteratively guide the self-training on a distributionally different target domain. To alleviate disruptions caused by the environmental discrepancy (e.g., beam numbers), the proposed cross-domain examination (CDE) assesses th

e correctness of pseudo labels by copy-pasting target instances into a source environment and measuring the prediction consistency. To reduce computational overhead and mitigate the object shift (e.g., scales and point densities), we design an overlapped boxes counting (OBC) metric that allows to uniformly downsample pseudo-labeled objects across different geometric characteristics. To confront the issue of inter-class imbalance, we progressively augment the target point clouds with a class-balanced set of pseudo-labeled target instances and source objects, which boosts recognition accuracies on both frequently appearing and rare classes. Experimental results on three benchmark datasets using both voxel-based (i.e., SECOND) and point-based 3D detectors (i.e., PointRCNN) demonstrate that our proposed ReDB approach outperforms existing 3D domain adaptation methods by a large margin, improving 23.15% mAP on the nuScenes - KITTI task. The code is available at https://github.com/zhuoxiao-chen/ReDB-DA-3Ddet.
****
Task-Oriented Multi-Modal Mutual Leaning for Vision-Language Models
Sifan Long, Zhen Zhao, Junkun Yuan, Zichang Tan, Jiangjiang Liu, Luping Zhou, Shengsheng Wang, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21959-21969
Prompt learning has become one of the most efficient paradigms for adapting large pre-trained vision-language models to downstream tasks. Current state-of-the-art methods, like CoOp and ProDA, tend to adopt soft prompts to learn an appropriate prompt for each specific task. Recent CoCoOp further boosts the base-to-new generalization performance via an image-conditional prompt. However, it directly fuses identical image semantics to prompts of different labels and significantly weakens the discrimination among different classes as shown in our experiments. Motivated by this observation, we first propose a class-aware text prompt (CTP) to enrich generated prompts with label-related image information. Unlike CoCoOp, CTP can effectively involve image semantics and avoid introducing extra ambiguities into different prompts. On the other hand, instead of reserving the complete image representations, we propose text-guided feature tuning (TFT) to make the image branch attend to class-related representation. A contrastive loss is employed to align such augmented text and image representations on downstream tasks. In this way, the image-to-text CTP and text-to-image TFT can be mutually promoted to enhance the adaptation of VLMs for downstream tasks. Extensive experiments demonstrate that our method outperforms the existing methods by a significant margin. Especially, compared to CoCoOp, we achieve an average improvement of 4.03% on new classes and 3.19% on harmonic-mean over eleven classification benchmarks.
****
Efficient Adaptive Human-Object Interaction Detection with Concept-guided Memory
Ting Lei, Fabian Caba, Qingchao Chen, Hailin Jin, Yuxin Peng, Yang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6480-6490
Human Object Interaction (HOI) detection aims to localize and infer the relationships between a human and an object. Arguably, training supervised models for this task from scratch presents challenges due to the performance drop over rare classes and the high computational cost and time required to handle long-tailed distributions of HOIs in complex HOI scenes in realistic settings. This observation motivates us to design an HOI detector that can be trained even with long-tailed labeled data and can leverage existing knowledge from pre-trained models. Inspired by the powerful generalization ability of the large Vision-Language Models (VLM) on classification and retrieval tasks, we propose an efficient Adaptive HOI Detector with Concept-guided Memory (ADA-CM). ADA-CM has two operating modes. The first mode makes it tunable without learning new parameters in a training-free paradigm. Its second mode incorporates an instance-aware adapter mechanism that can further efficiently boost performance if updating a lightweight set of parameters can be afforded. Our proposed method achieves competitive results with state-of-the-art on the HICO-DET and V-COCO datasets with much less training time. Code can be found at https://github.com/ltttpku/ADA-CM.
****

NeMF: Inverse Volume Rendering with Neural Microflake Field
Youjia Zhang, Teng Xu, Junqing Yu, Yuteng Ye, Yanqing Jing, Junle Wang, Jingyi Yu, Wei Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22919-22929
Recovering the physical attributes of an object's appearance from its images captured under an unknown illumination is challenging yet essential for photo-realistic rendering.Recent approaches adopt the emerging implicit scene representations and have shown impressive results.However, they unanimously adopt a surface-based representation,and hence can not well handle scenes with very complex geometry, translucent object and etc.In this paper, we propose to conduct inverse volume rendering, in contrast to surface-based, by representing a scene using microflake volume, which assumes the space is filled with infinite small flakes and light reflects or scatters at each spatial location according to microflake distributions. We further adopt the coordinate networks to implicitly encode the microflake volume, and develop a differentiable microflake volume renderer to train the network in an end-to-end way in principle.Our NeMF enables effective recovery of appearance attributes for highly complex geometry and scattering object, enables high-quality relighting, material editing, and especially simulates volume rendering effects, such as scattering, which is infeasible for surface-based approaches. Our data and code are available at: https://github.com/YoujiaZhang/NeMF.
****

Attentive Mask CLIP
Yifan Yang, Weiquan Huang, Yixuan Wei, Houwen Peng, Xinyang Jiang, Huiqiang Jiang, Fangyun Wei, Yin Wang, Han Hu, Lili Qiu, Yuqing Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2771-2781
In vision-language modeling, image token removal is an efficient augmentation technique to reduce the cost of encoding image features. The CLIP-style models, however, have been found to be negatively impacted by this technique. We hypothesize that removing a large portion of image tokens may inadvertently destroy the semantic information associated to a given text description, resulting in misaligned paired data in CLIP training. To address this issue, we propose an attentive token removal approach, which retains a small number of tokens that have a strong semantic correlation to the corresponding text description. The correlation scores are dynamically evaluated through an EMA-updated vision encoder. Our method, termed attentive mask CLIP, outperforms original CLIP and CLIP variant with random token removal while saving the training time. In addition, our approach also enables efficient multi-view contrastive learning. Experimentally, by training ViT-B on YFCC-15M dataset, our approach achieves 43.9% top-1 accuracy on ImageNet-1K zero-shot classification, 62.7/42.1 and 38.0/23.2 I2T/T2I retrieval accuracy on Flickr30K and MS COCO, outperforming SLIP by +1.1%,+5.5/+0.9, and +4.4/+1.3, respectively, while being 2.30x faster. An efficient version of our approach runs 1.16x faster than the plain CLIP model, while achieving significant gains of +5.3%, +11.3/+8.0, and +9.5/+4.9 on these benchmarks, respectively.
****

DOLCE: A Model-Based Probabilistic Diffusion Framework for Limited-Angle CT Reconstruction
Jiaming Liu, Rushil Anirudh, Jayaraman J. Thiagarajan, Stewart He, K Aditya Mohan, Ulugbek S. Kamilov, Hyojin Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10498-10508
Limited-Angle Computed Tomography (LACT) is a non-destructive 3D imaging technique used in a variety of applications ranging from security to medicine. The limited angle coverage in LACT is often a dominant source of severe artifacts in the reconstructed images, making it a challenging imaging inverse problem. Diffusion models are a recent class of deep generative models for synthesizing realistic images using image denoisers. In this work, we present DOLCE as the first framework for integrating conditionally-trained diffusion models and explicit physical measurement models for solving imaging inverse problems. DOLCE achieves the SOTA performance in highly ill-posed LACT by alternating between the data-fidelity and sampling updates of a diffusion model conditioned on the transformed sinogr

am. We show through extensive experimentation that unlike existing methods, DOLC
E can synthesize high-quality and structurally coherent 3D volumes by using only
 2D conditionally pre-trained diffusion models. We further show on several chall
enging real LACT datasets that the same pre-trained DOLCE model achieves the SOT
A performance on drastically different types of images.
****
Beyond Image Borders: Learning Feature Extrapolation for Unbounded Image Composi
tion
Xiaoyu Liu, Ming Liu, Junyi Li, Shuai Liu, Xiaotao Wang, Lei Lei, Wangmeng Zuo;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2023, pp. 13023-13032
For improving image composition and aesthetic quality, most existing methods mod
ulate the captured images by striking out redundant content near the image borde
rs. However, such image cropping methods are limited in the range of image views
. Some methods have been suggested to extrapolate the images and predict croppin
g boxes from the extrapolated image. Nonetheless, the synthesized extrapolated r
egions may be included in the cropped image, making the image composition result
 not real and potentially with degraded image quality. In this paper, we circumv
ent this issue by presenting a joint framework for both unbounded recommendation
 of camera view and image composition (i.e., UNIC). In this way, the cropped ima
ge is a sub-image of the image acquired by the predicted camera view, and thus c
an be guaranteed to be real and consistent in image quality. Specifically, our f
ramework takes the current camera preview frame as input and provides a recommen
dation for view adjustment, which contains operations unlimited by the image bor
ders, such as zooming in or out and camera movement. To improve the prediction a
ccuracy of view adjustment prediction, we further extend the field of view by fe
ature extrapolation. After one or several times of view adjustments, our method
converges and results in both a camera view and a bounding box showing the image
 composition recommendation. Extensive experiments are conducted on the datasets
 constructed upon existing image cropping datasets, showing the effectiveness of
 our UNIC in unbounded recommendation of camera view and image composition. The
source code, dataset, and pretrained models is available at https://github.com/l
iuxiaoyu1104/UNIC.
****
MasaCtrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthes
is and Editing
Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, Yinqiang Zheng; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 22560-22570
Despite the success in large-scale text-to-image generation and text-conditioned
 image editing, existing methods still struggle to produce consistent generation
 and editing results. For example, generation approaches usually fail to synthes
ize multiple images of the same objects/characters but with different views or p
oses. Meanwhile, existing editing methods either fail to achieve effective compl
ex non-rigid editing while maintaining the overall textures and identity, or req
uire time-consuming fine-tuning to capture the image-specific appearance. In thi
s paper, we develop MasaCtrl, a tuning-free method to achieve consistent image g
eneration and complex non-rigid image editing simultaneously. Specifically, Masa
Ctrl converts existing self-attention in diffusion models into mutual self-atten
tion, so that it can query correlated local contents and textures from source im
ages for consistency. To further alleviate the query confusion between foregroun
d and background, we propose a mask-guided mutual self-attention strategy, where
 the mask can be easily extracted from the cross-attention maps. Extensive exper
iments show that the proposed MasaCtrl can produce impressive results in both co
nsistent image generation and complex non-rigid real image editing.
****
Understanding Hessian Alignment for Domain Generalization
Sobhan Hemati, Guojun Zhang, Amir Estiri, Xi Chen; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 19004-19014
Out-of-distribution (OOD) generalization is a critical ability for deep learning

models in many real-world scenarios including healthcare and autonomous vehicles. Recently, different techniques have been proposed to improve OOD generalization. Among these methods, gradient-based regularizers have shown promising performance compared with other competitors. Despite this success, our understanding of the role of Hessian and gradient alignment in domain generalization is still limited. To address this shortcoming, we analyze the role of the classifier's head Hessian matrix and gradient in domain generalization using recent OoD theory of transferability. Theoretically, we show that spectral norm between the classifier's head Hessian matrices across domains is an upper bound of the transfer measure, a notion of distance between target and source domains. Furthermore, we analyze all the attributes that get aligned when we encourage similarity between Hessians and gradients. Our analysis explains the success of many regularizers like CORAL, IRM, V-REx, Fish, IGA, and Fishr as they regularize part of the classifier's head Hessian and/or gradient. Finally, we propose two simple yet effective methods to match the classifier's head Hessians and gradients in an efficient way, based on the Hessian Gradient Product (HGP) and Hutchinson's method (Hutchinson), and without directly calculating Hessians. We validate the OOD generalization ability of proposed methods in different scenarios, including transferability, severe correlation shift, label shift and diversity shift. Our results show that Hessian alignment methods achieve promising performance on various OOD benchmarks. Our code is available here.
****

DeepChange: A Long-Term Person Re-Identification Benchmark with Clothes Change
Peng Xu, Xiatian Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11196-11205
Long-term re-id with clothes change is a challenging problem in surveillance AI. Currently, its major bottleneck is that this field is still missing a large realistic benchmark. In this work, we contribute a large, realistic long-term person re-identification benchmark, termed DeepChange. Its unique characteristics are: (1) Realistic and rich personal appearance (e.g., clothes and hair style) and variations: Highly diverse clothes change and styles, with varying reappearing gaps in time from minutes to seasons, different weather conditions (e.g., sunny, cloudy, windy, rainy, snowy, extremely cold) and events (e.g., working, leisure, daily activities). (2) Rich camera setups: Raw videos were recorded by 17 outdoor varying-resolution cameras operating in a real-world surveillance system. (3) The currently largest number of (17) cameras, (1, 121) identities, and (178, 407) bounding boxes, over the longest time span (12 months). We benchmark the representative supervised and unsupervised re-id methods on our dataset. In addition, we investigate multimodal fusion strategies for tackling the clothes change challenge. Extensive experiments show that our fusion models outperform a wide variety of state-of-the-art models on DeepChange. Our dataset and documents are available at https://github.com/PengBoXiangShang/deepchange.
****

Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models
Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, Yogesh Balaji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22930-22941
Despite tremendous progress in generating high-quality images using diffusion models, synthesizing a sequence of animated frames that are both photorealistic and temporally coherent is still in its infancy. While off-the-shelf billion-scale datasets for image generation are available, collecting similar video data of the same scale is still challenging. Also, training a video diffusion model is computationally much more expensive than its image counterpart. In this work, we explore finetuning a pretrained image diffusion model with video data as a practical solution for the video synthesis task. We find that naively extending the image noise prior to video noise prior in video diffusion leads to sub-optimal performance. Our carefully designed video noise prior leads to substantially better performance. Extensive experimental validation shows that our model, Preserve Your Own COrrelation (PYoCo), attains SOTA zero-shot text-to-video results on the UCF-101 and MSR-VTT benchmarks. It also achieves SOTA video generation quality

on the small-scale UCF-101 benchmark with a 10x smaller model using significantl y less computation than the prior art.
****
Discrepant and Multi-Instance Proxies for Unsupervised Person Re-Identification
Chang Zou, Zeqi Chen, Zhichao Cui, Yuehu Liu, Chi Zhang; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2023, pp. 11058-11068
Most recent unsupervised person re-identification methods maintain a cluster uni -proxy for contrastive learning. However, due to the intra-class variance and in ter-class similarity, the cluster uni-proxy is prone to be biased and confused w ith similar classes, resulting in the learned features lacking intra-class compa ctness and inter-class separation in the embedding space. To completely and accu rately represent the information contained in a cluster and learn discriminative features, we propose to maintain discrepant cluster proxies and multi-instance proxies for a cluster. Each cluster proxy focuses on representing a part of the information, and several discrepant proxies collaborate to represent the entire cluster completely. As a complement to the overall representation, multi-instanc e proxies are used to accurately represent the fine-grained information containe d in the instances of the cluster. Based on the proposed discrepant cluster prox ies, we construct cluster contrastive loss to use the proxies as hard positive s amples to pull instances of a cluster closer and reduce intra-class variance. Me anwhile, instance contrastive loss is constructed by global hard negative sample mining in multi-instance proxies to push away the truly indistinguishable class es and decrease inter-class similarity. Extensive experiments on Market-1501 and MSMT17 demonstrate that the proposed method outperforms state-of-the-art approa ches.
****
Joint-Relation Transformer for Multi-Person Motion Prediction
Qingyao Xu, Weibo Mao, Jingze Gong, Chenxin Xu, Siheng Chen, Weidi Xie, Ya Zhang , Yanfeng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9816-9826
Multi-person motion prediction is a challenging problem due to the dependency of motion on both individual past movements and interactions with other people. Tr ansformer-based methods have shown promising resultson this task, but they miss the explicit relation representation between joints, such as skeleton structure and pairwise distance, which is crucial for accurate interaction modeling. In th is paper, we propose the Joint-Relation Transformer, which utilizes relation inf ormation to enhance interaction modeling and improve future motion prediction. O ur relation information contains the relative distance and the intra/inter-perso n physical constraints. To fuse relation and joint information, we design a nove l joint-relation fusion layer with relation-aware attention to update both featu res. Additionally, we supervise the relation information by forecasting future d istance. Experiments show that our method achieves a 13.4% improvement of 900ms VIM on 3DPW-SoMoF/RC and 17.8%/12.0% improvement of 3s MPJPE on CMU-Mpcap/MuPoTS -3D dataset.
****
Revisiting Vision Transformer from the View of Path Ensemble
Shuning Chang, Pichao Wang, Hao Luo, Fan Wang, Mike Zheng Shou; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19889- 19899
Vision Transformers (ViTs) are normally regarded as a stack of transformer layer s. In this work, we propose a novel view of ViTs showing that they can be seen a s ensemble networks containing multiple parallel paths with different lengths. S pecifically, we equivalently transform the traditional cascade of multi-head sel f-attention (MSA) and feed-forward network (FFN) into three parallel paths in ea ch transformer layer. Then, we utilize the identity connection in our new transf ormer form and further transform the ViT into an explicit multi-path ensemble ne twork. From the new perspective, these paths perform two functions: the first is to provide the feature for the classifier directly, and the second is to provid e the lower-level feature representation for subsequent longer paths. We investi gate the influence of each path for the final prediction and discover that some

paths even pull down the performance. Therefore, we propose the path pruning and EnsembleScale skills for improvement, which cut out the underperforming paths and re-weight the ensemble components, respectively, to optimize the path combination and make the short paths focus on providing high-quality representation for subsequent paths. We also demonstrate that our path combination strategies can help ViTs go deeper and act as high-pass filters to filter out partial low-frequency signals. To further enhance the representation of paths served for subsequent paths, self-distillation is applied to transfer knowledge from the long paths to the short paths. This work calls for more future research to explain and design ViTs from new perspectives.
****
Tetra-NeRF: Representing Neural Radiance Fields Using Tetrahedra
Jonas Kulhanek, Torsten Sattler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18458-18469
Neural Radiance Fields (NeRFs) are a very recent and very popular approach for the problems of novel view synthesis and 3D reconstruction. A popular scene representation used by NeRFs is to combine a uniform, voxel-based subdivision of the scene with an MLP. Based on the observation that a (sparse) point cloud of the scene is often available, this paper proposes to use an adaptive representation based on tetrahedra obtained by Delaunay triangulation instead of uniform subdivision or point-based representations. We show that such a representation enables efficient training and leads to state-of-the-art results. Our approach elegantly combines concepts from 3D geometry processing, triangle-based rendering, and modern neural radiance fields. Compared to voxel-based representations, ours provides more detail around parts of the scene likely to be close to the surface. Compared to point-based representations, our approach achieves better performance. The source code is publicly available at: https://jkulhanek.com/tetra-nerf.
****
TMA: Temporal Motion Aggregation for Event-based Optical Flow
Haotian Liu, Guang Chen, Sanqing Qu, Yanping Zhang, Zhijun Li, Alois Knoll, Changjun Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9685-9694
Event cameras have the ability to record continuous and detailed trajectories of objects with high temporal resolution, thereby providing intuitive motion cues for optical flow estimation. Nevertheless, most existing learning-based approaches for event optical flow estimation directly remould the paradigm of conventional images by representing the consecutive event stream as static frames, ignoring the inherent temporal continuity of event data. In this paper, we argue that temporal continuity is a vital element of event-based optical flow and propose a novel Temporal Motion Aggregation (TMA) approach to unlock its potential. Technically, TMA comprises three components: an event splitting strategy to incorporate intermediate motion information underlying the temporal context, a linear lookup strategy to align temporally fine-grained motion features and a novel motion pattern aggregation module to emphasize consistent patterns for motion feature enhancement. By incorporating temporally fine-grained motion information, TMA can derive better flow estimates than existing methods at early stages, which not only enables TMA to obtain more accurate final predictions, but also greatly reduces the demand for a number of refinements. Extensive experiments on DSEC-Flow and MVSEC datasets verify the effectiveness and superiority of our TMA. Remarkably, compared to E-RAFT, TMA achieves a 6% improvement in accuracy and a 40% reduction in inference time on DSEC-Flow. Code will be available at https://github.com/ispc-lab/TMA.
****
Ablating Concepts in Text-to-Image Diffusion Models
Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, Jun-Yan Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22691-22702
Large-scale text-to-image diffusion models can generate high-fidelity images with powerful compositional ability. However, these models are typically trained on an enormous amount of Internet data, often containing copyrighted material, lic

ensed images, and personal photos. Furthermore, they have been found to replicate the style of various living artists or memorize exact training samples. How can we remove such copyrighted concepts or images without retraining the model from scratch? To achieve this goal, we propose an efficient method of ablating concepts in the pretrained model, i.e., preventing the generation of a target concept. Our algorithm learns to match the image distribution for a target style, instance, or text prompt we wish to ablate to the distribution corresponding to an anchor concept. This prevents the model from generating target concepts given its text condition.

Extensive experiments show that our method can successfully prevent the generation of the ablated concept while preserving closely related concepts in the model.

****

Motion-Guided Masking for Spatiotemporal Representation Learning

David Fan, Jue Wang, Shuai Liao, Yi Zhu, Vimal Bhat, Hector Santos-Villalobos, Rohith MV, Xinyu Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5619-5629

Several recent works have directly extended the image masked autoencoder (MAE) with random masking into video domain, achieving promising results. However, unlike images, both spatial and temporal information are important for video understanding. This suggests that the random masking strategy that is inherited from the image MAE is less effective for video MAE. This motivates the design of a novel masking algorithm that can more efficiently make use of video saliency. Specifically, we propose a motion-guided masking algorithm (MGM) which leverages motion vectors to guide the position of each mask over time. Crucially, these motion-based correspondences can be directly obtained from information stored in the compressed format of the video, which makes our method efficient and scalable. On two challenging large-scale video benchmarks (Kinetics-400 and Something-Something V2), we equip video MAE with our MGM and achieve up to +1.3% improvement compared to previous state-of-the-art methods. Additionally, our MGM achieves equivalent performance to previous video MAE using up to 66% fewer training epochs. Lastly, we show that MGM generalizes better to downstream transfer learning and domain adaptation tasks on the UCF101, HMDB51, and Diving48 datasets, achieving up to +4.9% improvement compared to baseline methods.

****

MapFormer: Boosting Change Detection by Using Pre-change Information

Maximilian Bernhard, Niklas Strauß, Matthias Schubert; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16837-16846

Change detection in remote sensing imagery is essential for a variety of applications such as urban planning, disaster management, and climate research. However, existing methods for identifying semantically changed areas overlook the availability of semantic information in the form of existing maps describing features of the earth's surface. In this paper, we leverage this information for change detection in bi-temporal images. We show that the simple integration of the additional information via concatenation of latent representations suffices to significantly outperform state-of-the-art change detection methods. Motivated by this observation, we propose the new task of Conditional Change Detection, where pre-change semantic information is used as input next to bi-temporal images. To fully exploit the extra information, we propose MapFormer, a novel architecture based on a multi-modal feature fusion module that allows for feature processing conditioned on the available semantic information. We further employ a supervised, cross-modal contrastive loss to guide the learning of visual representations. Our approach outperforms existing change detection methods by an absolute 11.7% and 18.4% in terms of binary change IoU on DynamicEarthNet and HRSCD, respectively. Furthermore, we demonstrate the robustness of our approach to the quality of the pre-change semantic information and the absence pre-change imagery. The code is available at https://github.com/mxbh/mapformer.

****

Masked Diffusion Transformer is a Strong Image Synthesizer

Shanghua Gao, Pan Zhou, Ming-Ming Cheng, Shuicheng Yan; Proceedings of the IEEE/

CVF International Conference on Computer Vision (ICCV), 2023, pp. 23164-23173
Despite its success in image synthesis, we observe that diffusion probabilistic models (DPMs) often lack contextual reasoning ability to learn the relations among object parts in an image, leading to a slow learning process. To solve this issue, we propose a Masked Diffusion Transformer (MDT) that introduces a mask latent modeling scheme to explicitly enhance the DPMs' ability to contextual relation learning among object semantic parts in an image. During training, MDT operates in the latent space to mask certain tokens. Then, an asymmetric masking diffusion transformer is designed to predict masked tokens from unmasked ones while maintaining the diffusion generation process. Our MDT can reconstruct the full information of an image from its incomplete contextual input, thus enabling it to learn the associated relations among image tokens. Experimental results show that MDT achieves superior image synthesis performance, e.g., a new SOTA FID score in the ImageNet data set, and has about 3x faster learning speed than the previous SOTA DiT. The source code is released at https://github.com/sail-sg/MDT.
****
LightDepth: Single-View Depth Self-Supervision from Illumination Decline
Javier Rodríguez-Puigvert, Víctor M. Batlle, J.M.M. Montiel, Ruben Martinez-Cantin, Pascal Fua, Juan D. Tardós, Javier Civera; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21273-21283
Single-view depth estimation can be remarkably effective if there is enough ground-truth depth data for supervised training. However, there are scenarios, especially in medicine in the case of endoscopies, where such data cannot be obtained. In such cases, multi-view self-supervision and synthetic-to-real transfer serve as alternative approaches, however, with a considerable performance reduction in comparison to supervised case.
  Instead, we propose a single-view self-supervised method that achieves a performance similar to the supervised case. In some medical devices, such as endoscopes, the camera and light sources are co-located at a small distance from the target surfaces. Thus, we can exploit that, for any given albedo and surface orientation, pixel brightness is inversely proportional to the square of the distance to the surface, providing a strong single-view self-supervisory signal. In our experiments, our self-supervised models deliver accuracies comparable to those of fully supervised ones, while being applicable without depth ground-truth data.
****
Urban Radiance Field Representation with Deformable Neural Mesh Primitives
Fan Lu, Yan Xu, Guang Chen, Hongsheng Li, Kwan-Yee Lin, Changjun Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 465-476
Neural Radiance Fields (NeRFs) have achieved great success in the past few years. However, most current methods still require intensive resources due to ray marching-based rendering. To construct urban-level radiance fields efficiently, we design Deformable Neural Mesh Primitive (DNMP), and propose to parameterize the entire scene with such primitives. The DNMP is a flexible and compact neural variant of classic mesh representation, which enjoys both the efficiency of rasterization-based rendering and the powerful neural representation capability for photo-realistic image synthesis. Specifically, a DNMP consists of a set of connected deformable mesh vertices with paired vertex features to parameterize the geometry and radiance information of a local area. To constrain the degree of freedom for optimization and lower the storage budgets, we enforce the shape of each primitive to be decoded from a relatively low-dimensional latent space. The rendering colors are decoded from the vertex features (interpolated with rasterization) by a view-dependent MLP. The DNMP provides a new paradigm for urban-level scene representation with appealing properties: (1) High-quality rendering. Our method achieves leading performance for novel view synthesis in urban scenarios. (2) Low computational costs. Our representation enables fast rendering (2.07ms/1k pixels) and low peak memory usage (110MB/1k pixels). We also present a lightweight version that can run 33xfaster than vanilla NeRFs, and comparable to the highly-optimized Instant-NGP (0.61 vs 0.71ms/1k pixels).
****

Adaptive Frequency Filters As Efficient Global Token Mixers
Zhipeng Huang, Zhizheng Zhang, Cuiling Lan, Zheng-Jun Zha, Yan Lu, Baining Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6049-6059
Recent vision transformers, large-kernel CNNs and MLPs have attained remarkable successes in broad vision tasks thanks to their effective information fusion in the global scope. However, their efficient deployments, especially on mobile devices, still suffer from noteworthy challenges due to the heavy computational costs of self-attention mechanisms, large kernels, or fully connected layers. In this work, we apply conventional convolution theorem to deep learning for addressing this and reveal that adaptive frequency filters can serve as efficient global token mixer. With this insight, we propose Adaptive Frequency Filtering (AFF) token mixer. This neural operator transfers a latent representation to the frequency domain via a Fourier transform and performs semantic-adaptive frequency filtering via an elementwise multiplication, which mathematically equals to a token mixing operation in the original latent space with a dynamic convolution kernel as large as the spatial resolution of this latent representation. We take AFF token mixers as primary neural operators to build a lightweight neural network, dubbed AFFNet. Extensive experiments demonstrate the effectiveness of our proposed AFF token mixer and show that AFFNet achieve superior accuracy and efficiency trade-offs compared to other lightweight network designs on broad visual tasks, including visual recognition and dense prediction tasks.
****
Referring Image Segmentation Using Text Supervision
Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, Rynson Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22124-22134
Existing Referring Image Segmentation (RIS) methods typically require expensive pixel-level or box-level annotations for supervision. In this paper, we observe that the referring texts used in RIS already provide sufficient information to localize the target object. Hence, we propose a novel weakly-supervised RIS framework to formulate the target localization problem as a classification process to differentiate between positive and negative text expressions. While the referring text expressions for an image are used as positive expressions, the referring text expressions from other images can be used as negative expressions for this image. Our framework has three main novelties. First, we propose a bilateral prompt method to facilitate the classification process, by harmonizing the domain discrepancy between visual and linguistic features. Second, we propose a calibration method to reduce noisy background information and improve the correctness of the response maps for target object localization. Third, we propose a positive response map selection strategy to generate high-quality pseudo-labels from the enhanced response maps, for training a segmentation network for RIS inference. For evaluation, we propose a new metric to measure localization accuracy. Experiments on four benchmarks show that our framework achieves promising performances to existing fully-supervised RIS methods while outperforming state-of-the-art weakly-supervised methods adapted from related areas. Code is available at https://github.com/fawnliu/TRIS.
****
Zolly: Zoom Focal Length Correctly for Perspective-Distorted Human Mesh Reconstruction
Wenjia Wang, Yongtao Ge, Haiyi Mei, Zhongang Cai, Qingping Sun, Yanjun Wang, Chunhua Shen, Lei Yang, Taku Komura; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3925-3935
As it is hard to calibrate single-view RGB images in the wild, existing 3D human mesh reconstruction (3DHMR) methods either use a constant large focal length or estimate one based on the background environment context, which can not tackle the problem of the torso, limb, hand or face distortion caused by perspective camera projection when the camera is close to the human body. The naive focal length assumptions can harm this task with the incorrectly formulated projection matrices. To solve this, we propose Zolly, the first 3DHMR method focusing on persp

ective-distorted images. Our approach begins with analysing the reason for persp ective distortion, which we find is mainly caused by the relative location of th e human body to the camera center. We propose a new camera model and a novel 2D representation, termed distortion image, which describes the 2D dense distortion scale of the human body. We then estimate the distance from distortion scale fe atures rather than environment context features. Afterwards, We integrate the di stortion feature with image features to reconstruct the body mesh. To formulate the correct projection matrix and locate the human body position, we simultaneou sly use perspective and weak-perspective projection loss. Since existing dataset s could not handle this task, we propose the first synthetic dataset PDHuman and extend two real-world datasets tailored for this task, all containing perspecti ve-distorted human images. Extensive experiments show that Zolly outperforms exi sting state-of-the-art methods on both perspective-distorted datasets and the st andard benchmark (3DPW). Code and dataset will be released at https://wenjiawang 0312.github.io/projects/zolly/.
****
Once Detected, Never Lost: Surpassing Human Performance in Offline LiDAR based 3 D Object Detection
Lue Fan, Yuxue Yang, Yiming Mao, Feng Wang, Yuntao Chen, Naiyan Wang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19820-19829
This paper aims for high-performance offline LiDAR-based 3D object detection. We first observe that experienced human annotators annotate objects from a track-c entric perspective. They first label objects in a track with clear shapes, and t hen leverage the temporal coherence to infer the annotations of obscure objects. Drawing inspiration from this, we propose a high-performance offline detector i n a track-centric perspective instead of the conventional object-centric perspec tive. Our method features a bidirectional tracking module and a track-centric le arning module. Such a design allows our detector to infer and refine a complete track once the object is detected at a certain moment. We refer to this characte ristic as "onCe detecTed, neveR Lost" and name the proposed system CTRL. Extensi ve experiments demonstrate the remarkable performance of our method, surpassing the human-level annotating accuracy and outperforming the previous state-of-the- art methods in the highly competitive Waymo Open Dataset leaderboard without mod el ensemble. The code is available at https://github.com/tusen-ai/SST.
****
Building a Winning Team: Selecting Source Model Ensembles using a Submodular Tra nsferability Estimation Approach
Vimal K B, Saketh Bachu, Tanmay Garg, Niveditha Lakshmi Narasimhan, Raghavan Kon uru, Vineeth N Balasubramanian; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2023, pp. 11609-11620
Estimating the transferability of publicly available pre-trained models to a tar get task has assumed an important place for transfer learning tasks in recent ye ars. Existing efforts propose metrics that allow a user to choose one model from a pool of pre-trained models without having to fine-tune each model individuall y and identify one explicitly. With the growth in the number of available pre-tr ained models and the popularity of model ensembles, it also becomes essential to study the transferability of multiple-source models for a given target task. Th e few existing efforts study transferability in such multi-source ensemble setti ngs using just the outputs of the classification layer and neglect possible doma in or task mismatch. Moreover, they overlook the most important factor while sel ecting the source models, viz., the cohesiveness factor between them, which can impact the performance and confidence in the prediction of the ensemble. To addr ess these gaps, we propose a novel Optimal tranSport-based suBmOdular tRaNsferab ility metric (OSBORN) to estimate the transferability of an ensemble of models t o a downstream task. OSBORN collectively accounts for image domain difference, t ask difference, and cohesiveness of models in the ensemble to provide reliable e stimates of transferability. We gauge the performance of OSBORN on both image cl assification and semantic segmentation tasks. Our setup includes 28 source datas ets, 11 target datasets, 5 model architectures, and 2 pre-training methods. We b

enchmark our method against current state-of-the-art metrics MS-LEEP and E-LEEP, and outperform them consistently using the proposed approach.
****

Eventful Transformers: Leveraging Temporal Redundancy in Vision Transformers
Matthew Dutson, Yin Li, Mohit Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16911-16923
Vision Transformers achieve impressive accuracy across a range of visual recognition tasks. Unfortunately, their accuracy frequently comes with high computational costs. This is a particular issue in video recognition, where models are often applied repeatedly across frames or temporal chunks. In this work, we exploit temporal redundancy between subsequent inputs to reduce the cost of Transformers for video processing. We describe a method for identifying and re-processing only those tokens that have changed significantly over time. Our proposed family of models, Eventful Transformers, can be converted from existing Transformers (often without any re-training) and give adaptive control over the compute cost at runtime. We evaluate our method on large-scale datasets for video object detection (ImageNet VID) and action recognition (EPIC-Kitchens 100). Our approach leads to significant computational savings (on the order of 2-4x) with only minor reductions in accuracy.
****

Plausible Uncertainties for Human Pose Regression
Lennart Bramlage, Michelle Karg, Cristóbal Curio; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15133-15142
Human pose estimation (HPE) is integral to scene understanding in numerous safety-critical domains involving human-machine interaction, such as autonomous driving or semi-automated work environments. Avoiding costly mistakes is synonymous with anticipating failure in model predictions, which necessitates meta-judgments on the accuracy of the applied models. Here, we propose a straightforward human pose regression framework to examine the behavior of two established methods for simultaneous aleatoric and epistemic uncertainty estimation: maximum a-posteriori (MAP) estimation with Monte-Carlo variational inference and deep evidential regression (DER). First, we evaluate both approaches on the quality of their predicted variances and whether these truly capture the expected model error. The initial assessment indicates that both methods exhibit the overconfidence issue common in deep probabilistic models. This observation motivates our implementation of an additional recalibration step to extract reliable confidence intervals. We then take a closer look at deep evidential regression, which, to our knowledge, is applied comprehensively for the first time to the HPE problem. Experimental results indicate that DER behaves as expected in challenging and adverse conditions commonly occurring in HPE and that the predicted uncertainties match their purported aleatoric and epistemic sources. Notably, DER achieves smooth uncertainty estimates without the need for a costly sampling step, making it an attractive candidate for uncertainty estimation on resource-limited platforms.
****

Beyond One-to-One: Rethinking the Referring Image Segmentation
Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4067-4077
Referring image segmentation aims to segment the target object referred by a natural language expression. However, previous methods rely on the strong assumption that one sentence must describe one target in the image, which is often not the case in real-world applications. As a result, such methods fail when the expressions refer to either no objects or multiple objects. In this paper, we address this issue from two perspectives. First, we propose a Dual Multi-Modal Interaction (DMMI) Network, which contains two decoder branches and enables information flow in two directions. In the text-to-image decoder, text embedding is utilized to query the visual feature and localize the corresponding target. Meanwhile, the image-to-text decoder is implemented to reconstruct the erased entity-phrase conditioned on the visual feature. In this way, visual features are encouraged to contain the critical semantic information about target entity, which supports

the accurate segmentation in the text-to-image decoder in turn. Secondly, we collect a new challenging but realistic dataset called Ref-ZOM, which includes image-text pairs under different settings. Extensive experiments demonstrate our method achieves state-of-the-art performance on different datasets, and the Ref-ZOM-trained model performs well on various types of text inputs. Codes and datasets are available at https://github.com/toggle1995/RIS-DMMI.
****

Robust Referring Video Object Segmentation with Cyclic Structural Consensus
Xiang Li, Jinglu Wang, Xiaohao Xu, Xiao Li, Bhiksha Raj, Yan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22236-22245
Referring Video Object Segmentation (R-VOS) is a challenging task that aims to segment an object in a video based on a linguistic expression. Most existing R-VOS methods have a critical assumption: the object referred to must appear in the video. This assumption, which we refer to as "semantic consensus", is often violated in real-world scenarios, where the expression may be queried against false videos. In this work, we highlight the need for a robust R-VOS model that can handle semantic mismatches. Accordingly, we propose an extended task called Robust R-VOS (RRVOS), which accepts unpaired video-text inputs. We tackle this problem by jointly modeling the primary R-VOS problem and its dual (text reconstruction). A structural text-to-text cycle constraint is introduced to discriminate semantic consensus between video-text pairs and impose it in positive pairs, thereby achieving multi-modal alignment from both positive and negative pairs. Our structural constraint effectively addresses the challenge posed by linguistic diversity, overcoming the limitations of previous methods that relied on the point-wise constraint. A new evaluation dataset, RRYTVOS is constructed to measure the model robustness. Our model achieves state-of-the-art performance on R-VOS benchmarks, Ref-DAVIS17 and Ref-Youtube-VOS, and also our RRYTVOS dataset.
****

DiffIR: Efficient Diffusion Model for Image Restoration
Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13095-13105
Diffusion model (DM) has achieved SOTA performance by modeling the image synthesis process into a sequential application of a denoising network. However, different from image synthesis generating each pixel from scratch, most pixels of image restoration (IR) are given. Thus, for IR, traditional DMs running massive iterations on a large model to estimate whole images or feature maps is inefficient. To address this issue, we propose an efficient DM for IR (DiffIR), which consists of a compact IR prior extraction network (CPEN), dynamic IR transformer (DIRformer), and denoising network. Specifically, DiffIR has two training stages: pretraining and training DM. In pretraining, we input ground-truth images into CPEN-S1 to capture a compact IR prior representation (IPR) to guide DIRformer. In the second stage, we train the DM to directly estimate the same IRP as pretrained CPEN-S1 only using LQ images. We observe that since the IPR is only a compact vector, DiffIR can use fewer iterations than traditional DM to obtain accurate estimations and generate more stable and realistic results. Since the iterations are few, our DiffIR can adopt a joint optimization of CPEN-S2, DIRformer, and denoising network, which can further reduce the estimation error influence. We conduct extensive experiments on several IR tasks and achieve SOTA performance while consuming less computational costs. Codes and models will be released.
****

MoreauGrad: Sparse and Robust Interpretation of Neural Networks via Moreau Envelope
Jingwei Zhang, Farzan Farnia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2021-2030
Explaining the predictions of deep neural nets has been a topic of great interest in the computer vision literature. While several gradient-based interpretation schemes have been proposed to reveal the influential variables in a neural net's prediction, standard gradient-based interpretation frameworks have been common

ly observed to lack robustness to input perturbations and flexibility for incorporating prior knowledge of sparsity and group-sparsity structures. In this work, we propose MoreauGrad as an interpretation scheme based on the classifier neural net's Moreau envelope. We demonstrate that MoreauGrad results in a smooth and robust interpretation of a multi-layer neural network and can be efficiently computed through first-order optimization methods. Furthermore, we show that MoreauGrad can be naturally combined with L1-norm regularization techniques to output a sparse or group-sparse explanation which are prior conditions applicable to a wide range of deep learning applications. We empirically evaluate the proposed MoreauGrad scheme on standard computer vision datasets, showing the qualitative and quantitative success of the MoreauGrad approach in comparison to standard gradient-based interpretation methods.
****
Building Bridge Across the Time: Disruption and Restoration of Murals In the Wild
Huiyang Shao, Qianqian Xu, Peisong Wen, Peifeng Gao, Zhiyong Yang, Qingming Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20259-20269
In this paper, we focus on the mural-restoration task, which aims to detect damaged regions in the mural and repaint them automatically. Different from traditional image restoration tasks like in/out/blind-painting and image renovation, the corrupted mural suffers from more complicated degradation. However, existing mural-restoration methods and datasets still focus on simple degradation like masking. Such a significant gap prevents mural-restoration from being applied to real scenarios. To fill this gap, in this work, we propose a systematic framework to simulate the physical process for damaged murals and provide a new benchmark dataset for mural-restoration. Limited by the simplification of the data synthesis process, the previous mural-restoration methods suffer from poor performance in our proposed dataset. To handle this problem, we propose the Attention Diffusion Framework (ADF) for this challenging task. Within the framework, a damage attention map module is proposed to estimate the damage extent. Facing the diversity of defects, we propose a series of loss functions to choose repair strategies adaptively. Finally, experimental results support the effectiveness of the proposed framework in terms of both mural synthesis and restoration.
****
Class-Incremental Grouping Network for Continual Audio-Visual Learning
Shentong Mo, Weiguo Pian, Yapeng Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7788-7798
Continual learning is a challenging problem in which models need to be trained on non-stationary data across sequential tasks for class-incremental learning. While previous methods have focused on using either regularization or rehearsal-based frameworks to alleviate catastrophic forgetting in image classification, they are limited to a single modality and cannot learn compact class-aware cross-modal representations for continual audio-visual learning. To address this gap, we propose a novel class-incremental grouping network (CIGN) that can learn category-wise semantic features to achieve continual audio-visual learning. Our CIGN leverages learnable audio-visual class tokens and audio-visual grouping to continually aggregate class-aware features. Additionally, it utilizes class tokens distillation and continual grouping to prevent forgetting parameters learned from previous tasks, thereby improving the model's ability to capture discriminative audio-visual categories. We conduct extensive experiments on VGGSound-Instruments, VGGSound-100, and VGG-Sound Sources benchmarks. Our experimental results demonstrate that the CIGN achieves state-of-the-art audio-visual class-incremental learning performance.
****
Neural Haircut: Prior-Guided Strand-Based Hair Reconstruction
Vanessa Sklyarova, Jenya Chelishev, Andreea Dogaru, Igor Medvedev, Victor Lempitsky, Egor Zakharov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19762-19773
Generating realistic human 3D reconstructions using image or video data is essen

tial for various communication and entertainment applications. While existing me thods achieved impressive results for body and facial regions, realistic hair mo deling still remains challenging due to its high mechanical complexity. This wor k proposes an approach capable of accurate hair geometry reconstruction at a str and level from a monocular video or multi-view images captured in uncontrolled l ighting conditions. Our method has two stages, with the first stage performing j oint reconstruction of coarse hair and bust shapes and hair orientation using im plicit volumetric representations. The second stage then estimates a strand-leve l hair reconstruction by reconciling in a single optimization process the coarse volumetric constraints with hair strand and hairstyle priors learned from the s ynthetic data. To further increase the reconstruction fidelity, we incorporate i mage-based losses into the fitting process using a new differentiable renderer. The combined system, named Neural Haircut, achieves high realism and personaliza tion of the reconstructed hairstyles.
****
Improving Sample Quality of Diffusion Models Using Self-Attention Guidance
Susung Hong, Gyuseong Lee, Wooseok Jang, Seungryong Kim; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2023, pp. 7462-7471
Denoising diffusion models (DDMs) have attracted attention for their exceptional generation quality and diversity. This success is largely attributed to the use of class- or text-conditional diffusion guidance methods, such as classifier an d classifier-free guidance. In this paper, we present a more comprehensive persp ective that goes beyond the traditional guidance methods. From this generalized perspective, we introduce novel condition- and training-free strategies to enhan ce the quality of generated images. As a simple solution, blur guidance improves the suitability of intermediate samples for their fine-scale information and st ructures, enabling diffusion models to generate higher quality samples with a mo derate guidance scale. Improving upon this, Self-Attention Guidance (SAG) uses t he intermediate self-attention maps of diffusion models to enhance their stabili ty and efficacy. Specifically, SAG adversarially blurs only the regions that dif fusion models attend to at each iteration and guides them accordingly. Our exper imental results show that our SAG improves the performance of various diffusion models, including ADM, IDDPM, Stable Diffusion, and DiT. Moreover, combining SAG with conventional guidance methods leads to further improvement.
****
Evaluating Data Attribution for Text-to-Image Models
Sheng-Yu Wang, Alexei A. Efros, Jun-Yan Zhu, Richard Zhang; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7192-7203
While large text-to-image models are able to synthesize "novel" images, these im ages are necessarily a reflection of the training data. The problem of data attr ibution in such models -- which of the images in the training set are most respo nsible for the appearance of a given generated image -- is a difficult yet impor tant one. As an initial step toward this problem, we evaluate attribution throug h "customization" methods, which tune an existing large-scale model toward a giv en exemplar object or style. Our key insight is that this allows us to efficient ly create synthetic images that are computationally influenced by the exemplar b y construction. With our new dataset of such exemplar-influenced images, we are able to evaluate various data attribution algorithms and different possible feat ure spaces. Furthermore, by training on our dataset, we can tune standard models , such as DINO, CLIP, and ViT, toward the attribution problem. Even though the p rocedure is tuned towards small exemplar sets, we show generalization to larger sets. Finally, by taking into account the inherent uncertainty of the problem, w e can assign soft attribution scores over a set of training images.
****
Delta Denoising Score
Amir Hertz, Kfir Aberman, Daniel Cohen-Or,; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2023, pp. 2328-2337
This paper introduces Delta Denoising Score (DDS), a novel diffusion-based scori ng technique that optimizes a parametric model for the task of image editing. Un like the existing Score Distillation Sampling (SDS), which queries the generativ

e model with a single image-text pair, DDS utilizes an additional fixed query of a reference image-text pair to generate delta scores that represent the difference between the outputs of the two queries.
By estimating noisy gradient directions introduced by SDS using the source image and its text description, DDS provides cleaner gradient directions that modify the edited portions of the image while leaving others unchanged, yielding a distilled edit of the source image.
The analysis presented in this paper supports the power of the new score for image-to-image translation. We further show that the new score can be used to train an effective zero-shot image translation model.
The experimental results show that the proposed loss term outperforms existing methods in terms of stability and quality, highlighting its potential for real-world applications.
****
Hierarchical Prior Mining for Non-local Multi-View Stereo
Chunlin Ren, Qingshan Xu, Shikun Zhang, Jiaqi Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3611-3620
As a fundamental problem in computer vision, multi-view stereo (MVS) aims at recovering the 3D geometry of a target from a set of 2D images. Recent advances in MVS have shown that it is important to perceive non-local structured information for recovering geometry in low-textured areas. In this work, we propose a Hierarchical Prior Mining for Non-local Multi-View Stereo (HPM-MVS). The key characteristics are the following techniques that exploit non-local information to assist MVS: 1) A Non-local Extensible Sampling Pattern (NESP), which is able to adaptively change the size of sampled areas without becoming snared in locally optimal solutions. 2) A new approach to leverage non-local reliable points and construct a planar prior model based on K-Nearest Neighbor (KNN), to obtain potential hypotheses for the regions where prior construction is challenging. 3) A Hierarchical Prior Mining (HPM) framework, which is used to mine extensive non-local prior information at different scales to assist 3D model recovery, this strategy can achieve a considerable balance between the reconstruction of details and low-textured areas. Experimental results on the ETH3D and Tanks & Temples have verified the superior performance and strong generalization capability of our method. Our code will be available at https://github.com/CLinvx/HPM-MVS.
****
Generative Multiplane Neural Radiance for 3D-Aware Image Generation
Amandeep Kumar, Ankan Kumar Bhunia, Sanath Narayan, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Ming-Hsuan Yang, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7388-7398
We present a method to efficiently generate 3D-aware high-resolution images that are view-consistent across multiple target views. The proposed multiplane neural radiance model, named GMNR, consists of a novel a-guided view-dependent representation (a-VdR) module for learning view-dependent information. The a-VdR module, faciliated by an a-guided pixel sampling technique, computes the view-dependent representation efficiently by learning viewing direction and position coefficients. Moreover, we propose a view-consistency loss to enforce photometric similarity across multiple views. The GMNR model can generate 3D-aware high-resolution images that are view-consistent across multiple camera poses, while maintaining the computational efficiency in terms of both training and inference time. Experiments on three datasets demonstrate the effectiveness of the proposed modules, leading to favorable results in terms of both generation quality and inference time, compared to existing approaches. Our GMNR model generates 3D-aware images of 1024 x 1024 pixels with 17.6 FPS on a single V100. Code : https://github.com/VIROBO-15/GMNR
****
DG-Recon: Depth-Guided Neural 3D Scene Reconstruction
Jihong Ju, Ching Wei Tseng, Oleksandr Bailo, Georgi Dikov, Mohsen Ghafoorian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18184-18194
A key challenge in neural 3D scene reconstruction from monocular images is to fu

se features back projected from various views without any depth or occlusion information. We address this by leveraging monocular depth priors, which effectively guide the fusion to improve surface prediction and skip over irrelevant, ambiguous, or occluded features. Furthermore, we revisit the average-based fusion used by most neural 3D reconstruction methods and propose two alternatives, a variance-based and a cross-attention-based fusion module, that are more efficient and effective than the average-based and self-attention-based counterparts. Compared to the NeuralRecon baseline, the proposed DG-Recon models significantly improve the reconstruction quality and completeness while remaining in real-time. Our method achieves state-of-the-art online reconstruction results on the ScanNet dataset and is on par with the current best offline method, which repeatedly accesses keyframes from the entire video sequence. Our ScanNet-trained model also generalizes robustly to the challenging 7-Scenes dataset and a subset of SUN3D containing scenes as big as an entire floor.
****
Simple Baselines for Interactive Video Retrieval with Questions and Answers
Kaiqu Liang, Samuel Albanie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11091-11101
To date, the majority of video retrieval systems have been optimized for a "single-shot" scenario in which the user submits a query in isolation, ignoring previous interactions with the system. Recently, there has been renewed interest in interactive systems to enhance retrieval, but existing approaches are complex and deliver limited gains in performance. In this work, we revisit this topic and propose several simple yet effective baselines for interactive video retrieval via question-answering. We employ a VideoQA model to simulate user interactions and show that this enables the productive study of the interactive retrieval task without access to ground truth dialogue data. Experiments on MSR-VTT, MSVD, and AVSD show that our framework using question-based interaction significantly improves the performance of text-based video retrieval systems. Code is available at https://github.com/kevinliang888/IVR-QA-baselines.
****
MSRA-SR: Image Super-resolution Transformer with Multi-scale Shared Representation Acquisition
Xiaoqiang Zhou, Huaibo Huang, Ran He, Zilei Wang, Jie Hu, Tieniu Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12665-12676
Multi-scale feature extraction is crucial for many computer vision tasks, but it is rarely explored in Transformer-based image super-resolution (SR) methods. In this paper, we propose an image super-resolution Transformer with Multi-scale Shared Representation Acquisition (MSRA-SR). We incorporate the multi-scale feature acquisition into two basic Transformer modules, i.e., self-attention and feed-forward network. In particular, self-attention with cross-scale matching and convolution filters with different kernel sizes are designed to exploit the multi-scale features in images. Both global and multi-scale local features are explicitly extracted in the network. Moreover, we introduce a representation sharing mechanism to improve the efficiency of the multi-scale design. Analysis on the attention map correlation indicates the representation redundancy in self-attention, which motivates us to design a shared self-attention across different Transformer layers. The exhaustive element-wise similarity matching is computed only once and then shared by later layers. Besides, the multi-scale convolution in different branches can be equivalently transformed into a single convolution with reparameterization trick. Extensive experiments on lightweight, classical and real-world image SR tasks verify the effectiveness and efficiency of the proposed method.
****
The Stable Signature: Rooting Watermarks in Latent Diffusion Models
Pierre Fernandez, Guillaume Couairon, Hervé Jégou, Matthijs Douze, Teddy Furon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22466-22477
Generative image modeling enables a wide range of applications but raises ethica

l concerns about responsible deployment. This paper introduces an active strateg y combining image watermarking and Latent Diffusion Models. The goal is for all generated images to conceal a watermark allowing for future detection and/or ide ntification. The method quickly fine-tunes the image generator, conditioned on a binary signature. A pre-trained watermark extractor recovers the hidden signatu re from any generated image and a statistical test then determines whether it co mes from the generative model. We evaluate the invisibility and robustness of ou r watermark on a variety of generation tasks, showing that Stable Signature work s even after the images are modified. For instance, it detects the origin of an image generated from a text prompt, then cropped to keep 10% of the content, wit h 90+% accuracy at a false positive rate below 1e-6.
****
Boosting Semantic Segmentation from the Perspective of Explicit Class Embeddings
Yuhe Liu, Chuanjian Liu, Kai Han, Quan Tang, Zengchang Qin; Proceedings of the I EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 821-831
Semantic segmentation is a computer vision task that associates a label with eac h pixel in an image. Modern approaches tend to introduce class embeddings into s emantic segmentation for deeply utilizing category semantics, and regard supervi sed class masks as final predictions. In this paper, we explore the mechanism of class embeddings and have an insight that more explicit and meaningful class em beddings can be generated based on class masks purposely. Following this observa tion, we propose ECENet, a new segmentation paradigm, in which class embeddings are obtained and enhanced explicitly during interacting with multi-stage image f eatures. Based on this, we revisit the traditional decoding process and explore inverted information flow between segmentation masks and class embeddings. Furth ermore, to ensure the discriminability and informativity of features from backbo ne, we propose a Feature Reconstruction module, which combines intrinsic and div erse branches together to ensure the concurrence of diversity and redundancy in features. Experiments show that our ECENet outperforms its counterparts on the A DE20K dataset with much less computational cost and achieves new state-of-the-ar t results on PASCAL-Context dataset. The code will be released at https://gitee. com/mindspore/models and https://github.com/Carol-lyh/ECENet.
****
Going Denser with Open-Vocabulary Part Segmentation
Peize Sun, Shoufa Chen, Chenchen Zhu, Fanyi Xiao, Ping Luo, Saining Xie, Zhichen g Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15453-15465
Object detection has been expanded from a limited number of categories to open v ocabulary. Moving forward, a complete intelligent vision system requires underst anding more fine-grained object descriptions, object parts. In this paper, we pr opose a detector with the ability to predict both open-vocabulary objects and th eir part segmentation. This ability comes from two designs. First, we train the detector on the joint of part-level, object-level and image-level data to build the multi-granularity alignment between language and image. Second, we parse the novel object into its parts by its dense semantic correspondence with the base object. These two designs enable the detector to largely benefit from various da ta sources and foundation models. In open-vocabulary part segmentation experimen ts, our method outperforms the baseline by 3.3 7.3 mAP in cross-dataset generali zation on PartImageNet, and improves the baseline by 7.3 novel AP50 in cross-cat egory generalization on Pascal Part. Finally, we train a detector that generaliz es to a wide range of part segmentation datasets while achieving better performa nce than dataset-specific training.
****
Learning to Identify Critical States for Reinforcement Learning from Videos
Haozhe Liu, Mingchen Zhuge, Bing Li, Yuhui Wang, Francesco Faccio, Bernard Ghane m, Jürgen Schmidhuber; Proceedings of the IEEE/CVF International Conference on C omputer Vision (ICCV), 2023, pp. 1955-1965
Recent work on deep reinforcement learning (DRL) has pointed out that algorithmi c information about good policies can be extracted from offline data which lack explicit information about executed actions. For example, videos of humans or ro

bots may convey a lot of implicit information about rewarding action sequences, but a DRL machine that wants to profit from watching such videos must first learn by itself to identify and recognize relevant states/actions/rewards. Without relying on ground-truth annotations, our new method called Deep State Identifier learns to predict returns from episodes encoded as videos. Then it uses a kind of mask-based sensitivity analysis to extract/identify important critical states. Extensive experiments showcase our method's potential for understanding and improving agent behavior. The source code and the generated datasets are available at https://github.com/AI-Initiative-KAUST/VideoRLCS.
****
Editing Implicit Assumptions in Text-to-Image Diffusion Models
Hadas Orgad, Bahjat Kawar, Yonatan Belinkov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7053-7061
Text-to-image diffusion models often make implicit assumptions about the world when generating images.
While some assumptions are useful (e.g., the sky is blue), they can also be outdated, incorrect, or reflective of social biases present in the training data.
Thus, there is a need to control these assumptions without requiring explicit user input or costly re-training.
In this work, we aim to edit a given implicit assumption in a pre-trained diffusion model.
Our Text-to-Image Model Editing method, TIME for short, receives a pair of inputs: a "source" under-specified prompt for which the model makes an implicit assumption (e.g., "a pack of roses"), and a "destination" prompt that describes the same setting, but with a specified desired attribute (e.g., "a pack of blue roses").
TIME then updates the model's cross-attention layers, as these layers assign visual meaning to textual tokens.
We edit the projection matrices in these layers such that the source prompt is projected close to the destination prompt.
Our method is highly efficient, as it modifies a mere 2.2% of the model's parameters in under one second.
To evaluate model editing approaches, we introduce TIMED (TIME Dataset), containing 147 source and destination prompt pairs from various domains.
Our experiments (using Stable Diffusion) show that TIME is successful in model editing, generalizes well for related prompts unseen during editing, and imposes minimal effect on unrelated generations.
****
OCHID-Fi: Occlusion-Robust Hand Pose Estimation in 3D via RF-Vision
Shujie Zhang, Tianyue Zheng, Zhe Chen, Jingzhi Hu, Abdelwahed Khamis, Jiajun Liu, Jun Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15112-15121
Hand Pose Estimation (HPE) is crucial to many applications, but conventional cameras-based CM-HPE methods are completely subject to Line-of-Sight (LoS), as cameras cannot capture occluded objects. In this paper, we propose to exploit Radio-Frequency-Vision (RF-vision) capable of bypassing obstacles for achieving occluded HPE, and we introduce OCHID-Fi as the first RF-HPE method with 3D pose estimation capability. OCHID-Fi employs wideband RF sensors widely available on smart devices (e.g., iPhones) to probe 3D human hand pose and extract their skeletons behind obstacles. To overcome the challenge in labeling RF imaging given its human incomprehensible nature, OCHID-Fi employs a cross-modality and cross-domain training process. It uses a pre-trained CM-HPE network and a synchronized CM/RF dataset, to guide the training of its complex-valued RF-HPE network under LoS conditions. It further transfers knowledge learned from labeled LoS domain to unlabeled occluded domain via adversarial learning, enabling OCHID-Fi to generalize to unseen occluded scenarios. Experimental results demonstrate the superiority of OCHID-Fi: it achieves comparable accuracy to CM-HPE under normal conditions while maintaining such accuracy even in occluded scenarios, with empirical evidence for its generalizability to new domains.
****

Conceptual and Hierarchical Latent Space Decomposition for Face Editing
Savas Ozkan, Mete Ozay, Tom Robinson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7248-7257
Generative Adversarial Networks (GANs) can produce photo-realistic results using an unconditional image-generation pipeline. However, the images generated by GANs (e.g., StyleGAN) are entangled in feature spaces, which makes it difficult to interpret and control the contents of images. In this paper, we present an encoder-decoder model that decomposes the entangled GAN space into a conceptual and hierarchical latent space in a self-supervised manner. The outputs of 3D morphable face models are leveraged to independently control image synthesis parameters like pose, expression, and illumination. For this purpose, a novel latent space decomposition pipeline is introduced using transformer networks and generative models. Later, this new space is used to optimize a transformer-based GAN space controller for face editing. In this work, a StyleGAN2 model for faces is utilized. Since our method manipulates only GAN features, the photo-realism of StyleGAN2 is fully preserved. The results demonstrate that our method qualitatively and quantitatively outperforms baselines in terms of identity preservation and editing precision.
****
VL-Match: Enhancing Vision-Language Pretraining with Token-Level and Instance-Level Matching
Junyu Bi, Daixuan Cheng, Ping Yao, Bochen Pang, Yuefeng Zhan, Chuanguang Yang, Yujing Wang, Hao Sun, Weiwei Deng, Qi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2584-2593
Vision-Language Pretraining (VLP) has significantly improved the performance of various vision-language tasks with the matching of images and texts. In this paper, we propose VL-Match, a Vision-Language framework with Enhanced Token-level and Instance-level Matching. At the token level, a Vision-Language Replaced Token Detection task is designed to boost the substantial interaction between text tokens and images, where the text encoder of VLP works as a generator to generate a corrupted text, and the multimodal encoder of VLP works as a discriminator to predict whether each text token in the corrupted text matches the image. At the instance level, in the Image-Text Matching task that judges whether an image-text pair is matched, we propose a novel bootstrapping method to generate hard negative text samples that are different from the positive ones only at the token level. In this way, we can force the network to detect fine-grained differences between images and texts. Notably, with a smaller amount of parameters, VL-Match significantly outperforms previous SOTA on all image-text retrieval tasks.
****
Reconstructing Interacting Hands with Interaction Prior from Monocular Images
Binghui Zuo, Zimeng Zhao, Wenqian Sun, Wei Xie, Zhou Xue, Yangang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9054-9064
Reconstructing interacting hands from monocular images is indispensable in AR/VR applications. Most existing solutions rely on the accurate localization of each skeleton joint. However, these methods tend to be unreliable due to the severe occlusion and confusing similarity among adjacent hand parts. This also defies human perception because humans can quickly imitate an interaction pattern without localizing all joints. Our key idea is to first construct a two-hand interaction prior and recast the interaction reconstruction task as the conditional sampling from the prior. To expand more interaction states, a large-scale multimodal dataset with physical plausibility is proposed. Then a VAE is trained to further condense these interaction patterns as latent codes in a prior distribution. When looking for image cues that contribute to interaction prior sampling, we propose the interaction adjacency heatmap (IAH). Compared with a joint-wise heatmap for localization, IAH assigns denser visible features to those invisible joints. Compared with an all-in-one visible heatmap, it provides more fine-grained local interaction information in each interaction region. Finally, the correlations between the extracted features and corresponding interaction codes are linked by the ViT module. Comprehensive evaluations on benchmark datasets have verified t

he effectiveness of this framework. The code and dataset are publicly available at: https://github.com/binghui-z/InterPrior_pytorch.
****
Towards Realistic Evaluation of Industrial Continual Learning Scenarios with an Emphasis on Energy Consumption and Computational Footprint
Vivek Chavan, Paul Koch, Marian Schlüter, Clemens Briese; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11506-11518
Incremental Learning (IL) aims to develop Machine Learning (ML) models that can learn from continuous streams of data and mitigate catastrophic forgetting. We analyse the current state-of-the-art Class-IL implementations and demonstrate why the current body of research tends to be one-dimensional, with an excessive focus on accuracy metrics. A realistic evaluation of Continual Learning methods should also emphasise energy consumption and overall computational load for a comprehensive understanding. This paper addresses research gaps between current IL research and industrial project environments, including varying incremental tasks and the introduction of Joint Training in tandem with IL. We introduce InVar-100 (Industrial Objects in Varied Contexts), a novel dataset meant to simulate the visual environments in industrial setups and perform various experiments for IL. Additionally, we incorporate explainability (using class activations) to interpret the model predictions. Our approach, RECIL (Real-World Scenarios and Energy Efficiency Considerations for Class Incremental Learning) provides meaningful insights about the applicability of IL approaches in practical use cases. The overarching aim is to bring the Incremental Learning and Green AI fields together and encourage the application of CIL methods in real-world scenarios. Code and dataset are available.
****
Translating Images to Road Network: A Non-Autoregressive Sequence-to-Sequence Approach
Jiachen Lu, Renyuan Peng, Xinyue Cai, Hang Xu, Hongyang Li, Feng Wen, Wei Zhang, Li Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23-33
The extraction of road network is essential for the generation of high-definition maps since it enables the precise localization of road landmarks and their interconnections. However, generating road network poses a significant challenge due to the conflicting underlying combination of Euclidean (e.g., road landmarks location) and non-Euclidean (e.g., road topological connectivity) structures. Existing methods struggle to merge the two types of data domains effectively, but few of them address it properly. Instead, our work establishes a unified representation of both types of data domain by projecting both Euclidean and non-Euclidean data into an integer series called RoadNet Sequence. Further than modeling an auto-regressive sequence-to-sequence Transformer model to understand RoadNet Sequence, we decouple the dependency of RoadNet Sequence into a mixture of auto-regressive and non-autoregressive dependency. Building on this, our proposed non-autoregressive sequence-to-sequence approach leverages non-autoregressive dependencies while fixing the gap towards auto-regressive dependencies, resulting in success on both efficiency and accuracy. Extensive experiments on nuScenes dataset demonstrate the superiority of RoadNet Sequence representation and the non-autoregressive approach compared to existing state-of-the-art alternatives.
****
How Much Temporal Long-Term Context is Needed for Action Segmentation?
Emad Bahrami, Gianpiero Francesca, Juergen Gall; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10351-10361
Modeling long-term context in videos is crucial for many fine-grained tasks including temporal action segmentation. An interesting question that is still open is how much long-term temporal context is needed for optimal performance. While transformers can model the long-term context of a video, this becomes computationally prohibitive for long videos. Recent works on temporal action segmentation thus combine temporal convolutional networks with self-attentions that are computed only for a local temporal window. While these approaches show good results, their performance is limited by their inability to capture the full context of a

video. In this work, we try to answer how much long-term temporal context is required for temporal action segmentation by introducing a transformer-based model that leverages sparse attention to capture the full context of a video. We compare our model with the current state of the art on three datasets for temporal action segmentation, namely 50Salads, Breakfast, and Assembly101. Our experiments show that modeling the full context of a video is necessary to obtain the best performance for temporal action segmentation.
****
3D VR Sketch Guided 3D Shape Prototyping and Exploration
Ling Luo, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song, Yulia Gryaditskaya; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9267-9276
3D shape modeling is labor-intensive, time-consuming, and requires years of expertise. To facilitate 3D shape modeling, we propose a 3D shape generation network that takes a 3D VR sketch as a condition. We assume that sketches are created by novices without art training and aim to reconstruct geometrically realistic 3D shapes of a given category. To handle potential sketch ambiguity, our method creates multiple 3D shapes that align with the original sketch's structure. We carefully design our method, training the model step-by-step and leveraging multi-modal 3D shape representation to support training with limited training data. To guarantee the realism of generated 3D shapes we leverage the normalizing flow that models the distribution of the latent space of 3D shapes. To encourage the fidelity of the generated 3D shapes to an input sketch, we propose a dedicated loss that we deploy at different stages of the training process. The code is available at https://github.com/Rowl1ng/3Dsketch2shape.
****
Seal-3D: Interactive Pixel-Level Editing for Neural Radiance Fields
Xiangyu Wang, Jingsen Zhu, Qi Ye, Yuchi Huo, Yunlong Ran, Zhihua Zhong, Jiming Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17683-17693
With the popularity of implicit neural representations, or neural radiance fields (NeRF), there is a pressing need for editing methods to interact with the implicit 3D models for tasks like post-processing reconstructed scenes and 3D content creation. While previous works have explored NeRF editing from various perspectives, they are restricted in editing flexibility, quality, and speed, failing to offer direct editing response and instant preview. The key challenge is to conceive a locally editable neural representation that can directly reflect the editing instructions and update instantly. To bridge the gap, we propose a new interactive editing method and system for implicit representations, called Seal-3D, which allows users to edit NeRF models in a pixel-level and free manner with a wide range of NeRF-like backbone and preview the editing effects instantly. To achieve the effects, the challenges are addressed by our proposed proxy function mapping the editing instructions to the original space of NeRF models in the teacher model and a two-stage training strategy for the student model with local pretraining and global finetuning. A NeRF editing system is built to showcase various editing types. Our system can achieve compelling editing effects with an interactive speed of about 1 second.
****
Generative Novel View Synthesis with 3D-Aware Diffusion Models
Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, Gordon Wetzstein; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4217-4229
We present a diffusion-based model for 3D-aware generative novel view synthesis from as few as a single input image. Our model samples from the distribution of possible renderings consistent with the input and, even in the presence of ambiguity, is capable of rendering diverse and plausible novel views. To achieve this, our method makes use of existing 2D diffusion backbones but, crucially, incorporates geometry priors in the form of a 3D feature volume. This latent feature field captures the distribution over possible scene representations and improves

our method's ability to generate view-consistent novel renderings. In addition to generating novel views, our method has the ability to autoregressively synthesize 3D-consistent sequences. We demonstrate state-of-the-art results on synthetic renderings and room-scale scenes; we also show compelling results for challenging, real-world objects.
****

MDCS: More Diverse Experts with Consistency Self-distillation for Long-tailed Recognition
Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, Jun Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11597-11608
Recently, multi-expert methods have led to significant improvements in long-tail recognition (LTR). We summarize two aspects that need further enhancement to contribute to LTR boosting: (1) More diverse experts; (2) Lower model variance. However, the previous methods didn't handle them well. To this end, we propose More Diverse experts with Consistency Self-distillation (MDCS) to bridge the gap left by earlier methods. Our MDCS approach consists of two core components: Diversity Loss (DL) and Consistency Self-distillation (CS). In detail, DL promotes diversity among experts by controlling their focus on different categories. To reduce the model variance, we employ KL divergence to distill the richer knowledge of weakly augmented instances for the experts' self-distillation. In particular, we design Confident Instance Sampling (CIS) to select the correctly classified instances for CS to avoid biased/noisy knowledge. In the analysis and ablation study, we demonstrate that our method compared with previous work can effectively increase the diversity of experts, significantly reduce the variance of the model, and improve recognition accuracy. Moreover, the roles of our DL and CS are mutually reinforcing and coupled: the diversity of experts benefits from the CS, and the CS cannot achieve remarkable results without the DL. Experiments show our MDCS outperforms the state-of-the-art by 1%   2% on five popular long-tailed benchmarks, including CIFAR10-LT, CIFAR100-LT, ImageNet-LT, Places-LT, and iNaturalist 2018. The code is available at https://github.com/fistyee/MDCS.
****

Similarity Min-Max: Zero-Shot Day-Night Domain Adaptation
Rundong Luo, Wenjing Wang, Wenhan Yang, Jiaying Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8104-8114
Low-light conditions not only hamper human visual experience but also degrade the model's performance on downstream vision tasks. While existing works make remarkable progress on day-night domain adaptation, they rely heavily on domain knowledge derived from the task-specific nighttime dataset. This paper challenges a more complicated scenario with border applicability, i.e., zero-shot day-night domain adaptation, which eliminates reliance on any nighttime data. Unlike prior zero-shot adaptation approaches emphasizing either image-level translation or model-level adaptation, we propose a similarity min-max paradigm that considers them under a unified framework. On the image level, we darken images towards minimum feature similarity to enlarge the domain gap. Then on the model level, we maximize the feature similarity between the darkened images and their normal-light counterparts for better model adaptation. To the best of our knowledge, this work represents the pioneering effort in jointly optimizing both aspects, resulting in a significant improvement of model generalizability. Extensive experiments demonstrate our method's effectiveness and broad applicability on various nighttime vision tasks, including classification, semantic segmentation, visual place recognition, and video action recognition. Our project page is available at https://red-fairy.github.io/ZeroShotDayNightDA-Webpage/.
****

Dark Side Augmentation: Generating Diverse Night Examples for Metric Learning
Albert Mohwald, Tomas Jenicek, Ond■ej Chum; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11153-11163
Image retrieval methods based on CNN descriptors rely on metric learning from a large number of diverse examples of positive and negative image pairs. Domains, such as night-time images, with limited availability and variability of training data suffer from poor retrieval performance even with methods performing well o

n standard benchmarks. We propose to train a GAN-based synthetic-image generator
, translating available day-time image examples into night images. Such a genera
tor is used in metric learning as a form of augmentation, supplying training dat
a to the scarce domain. Various types of generators are evaluated and analyzed.
We contribute with a novel light-weight GAN architecture that enforces the consi
stency between the original and translated image through edge consistency. The p
roposed architecture also allows a simultaneous training of an edge detector tha
t operates on both night and day images. To further increase the variability in
the training examples and to maximize the generalization of the trained model, w
e propose a novel method of diverse anchor mining.

The proposed method improves over the state-of-the-art results on a standard To
kyo 24/7 day-night retrieval benchmark while preserving the performance on Oxfor
d and Paris datasets. This is achieved without the need of training image pairs
of matching day and night images. The source code is available at https://github
.com/mohwald/gandtr .
****
NeRF-MS: Neural Radiance Fields with Multi-Sequence
Peihao Li, Shaohui Wang, Chen Yang, Bingbing Liu, Weichao Qiu, Haoqian Wang; Pro
ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202
3, pp. 18591-18600
Neural radiance fields (NeRF) achieve impressive performance in novel view synth
esis when trained on only single sequence data. However, leveraging multiple seq
uences captured by different cameras at different times is essential for better
reconstruction performance. Multi-sequence data takes two main challenges: appea
rance variation due to different lighting conditions and non-static objects like
 pedestrians. To address these issues, we propose NeRF-MS, a novel approach to t
raining NeRF with multi-sequence data. Specifically, we utilize a triplet loss t
o regularize the distribution of per-image appearance code, which leads to bette
r high-frequency texture and consistent appearance, such as specular reflections
. Then, we explicitly model non-static objects to reduce floaters. Extensive res
ults demonstrate that NeRF-MS not only outperforms state-of-the-art view synthes
is methods on outdoor and synthetic scenes, but also achieves 3D consistent rend
ering and robust appearance controlling. Project page: https://nerf-ms.github.io
/.
****
LVOS: A Benchmark for Long-term Video Object Segmentation
Lingyi Hong, Wenchao Chen, Zhongying Liu, Wei Zhang, Pinxue Guo, Zhaoyu Chen, We
nqiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer V
ision (ICCV), 2023, pp. 13480-13492
Existing video object segmentation (VOS) benchmarks focus on short-term videos w
hich just last about 3-5 seconds and where objects are visible most of the time.
 These videos are poorly representative of practical applications, and the absen
ce of long-term datasets restricts further investigation of VOS on the applicati
on in realistic scenarios. So, in this paper, we present a new benchmark dataset
 and evaluation methodology named LVOS, which consists of 220 videos with a tota
l duration of 421 minutes. To the best of our knowledge, LVOS is the first dense
ly annotated long-term VOS dataset. The videos in our LVOS last 1.59 minutes on
average, which is 20 times longer than videos in existing VOS datasets. Each vid
eo includes various attributes, especially challenges deriving from the wild, su
ch as long-term reappearing and cross-temporal similar objects. Based on LVOS, w
e assess existing video object segmentation algorithms and propose a Diverse Dyn
amic Memory network (DDMemory) that consists of three complementary memory banks
 to exploit temporal information adequately. The experimental results demonstrat
e the strength and weaknesses of prior methods, pointing promising directions fo
r further study. Our objective is to provide the community with a large and vari
ed benchmark to boost the advancement of long-term VOS. Data and code are availa
ble at https://lingyihongfd.github.io/lvos.github.io/.
****
Diffusion Model as Representation Learner

Xingyi Yang, Xinchao Wang; Proceedings of the IEEE/CVF International Conference
on Computer Vision (ICCV), 2023, pp. 18938-18949
Diffusion Probabilistic Models (DPMs) have recently demonstrated impressive resu
lts on various generative tasks. Despite its promises, the learned representatio
ns of pre-trained DPMs, however, have not been fully understood.
 In this paper, we conduct an in-depth investigation of the representation power
 of DPMs, and propose a novel knowledge transfer method that leverages the knowl
edge acquired by generative DPMs for analytical tasks. Our study begins by exami
ning the feature space of DPMs, revealing that DPMs are inherently denoising aut
oencoders that balance the representation learning with regularizing model capac
ity. To this end, we introduce a novel knowledge transfer paradigm named RepFusi
on. Our paradigm extracts representations at different time steps from off-the-s
helf DPMs and dynamically employs them as supervision for student networks, in w
hich the optimal time is determined through reinforcement learning. We evaluate
our approach on several image classification, semantic segmentation, and landmar
k detection benchmarks, and demonstrate that it outperforms state-of-the-art met
hods. Our results uncover the potential of DPMs as a powerful tool for represent
ation learning and provide insights into the usefulness of generative models bey
ond sample generation.
****
Nerfbusters: Removing Ghostly Artifacts from Casually Captured NeRFs
Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, Angjoo Kanaz
awa; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 18120-18130
Casually captured Neural Radiance Fields (NeRFs) suffer from artifacts such as f
loaters or flawed geometry when rendered outside the input camera trajectory. Ex
isting evaluation protocols often do not capture these effects, since they usual
ly only assess image quality at every 8th frame of the training capture. To aid
in the development and evaluation of new methods in novel-view synthesis, we pro
pose a new dataset and evaluation procedure, where two camera trajectories are r
ecorded of the scene: one used for training, and the other for evaluation. In th
is more challenging in-the-wild setting, we find that existing hand-crafted regu
larizers do not remove floaters nor improve scene geometry. Thus, we propose a 3
D diffusion-based method that leverages local 3D priors and a novel density-base
d score distillation sampling loss to discourage artifacts during NeRF optimizat
ion. We show that this data-driven prior removes floaters and improves scene geo
metry for casual captures.
****
Document Understanding Dataset and Evaluation (DUDE)
Jordy Van Landeghem, Rubèn Tito, ■ukasz Borchmann, Micha■ Pietruszka, Pawel Jozi
ak, Rafal Powalski, Dawid Jurkiewicz, Mickael Coustaty, Bertrand Anckaert, Ernes
t Valveny, Matthew Blaschko, Sien Moens, Tomasz Stanislawek; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19528-195
40
We call on the Document AI (DocAI) community to re-evaluate current methodologie
s and embrace the challenge of creating more practically-oriented benchmarks. Do
cument Understanding Dataset and Evaluation (DUDE) seeks to remediate the halted
 research progress in understanding visually-rich documents (VRDs). We present a
 new dataset with novelties related to types of questions, answers, and document
 layouts based on multi-industry, multi-domain, and multi-page VRDs of various o
rigins and dates. Moreover, we are pushing the boundaries of current methods by
creating multi-task and multi-domain evaluation setups that more accurately simu
late real-world situations where powerful generalization and adaptation under lo
w-resource settings are desired. DUDE aims to set a new standard as a more pract
ical, long-standing benchmark for the community, and we hope that it will lead t
o future extensions and contributions that address real-world challenges. Finall
y, our work illustrates the importance of finding more efficient ways to model l
anguage, images, and layout in DocAI.
****
ALWOD: Active Learning for Weakly-Supervised Object Detection

Yuting Wang, Velibor Ilic, Jiatong Li, Branislav Kisa■anin, Vladimir Pavlovic; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6459-6469

Object detection (OD), a crucial vision task, remains challenged by the lack of large training datasets with precise object localization labels. In this work, we propose ALWOD, a new framework that addresses this problem by fusing active learning (AL) with weakly and semi-supervised object detection paradigms. Because the performance of AL critically depends on the model initialization, we propose a new auxiliary image generator strategy that utilizes an extremely small labeled set, coupled with a large weakly tagged set of images, as a warm-start for AL. We then propose a new AL acquisition function, another critical factor in AL success, that leverages the student-teacher OD pair disagreement and uncertainty to effectively propose the most informative images to annotate. Finally, to complete the AL loop, we introduce a new labeling task delegated to human annotators, based on selection and correction of model-proposed detections, which is both rapid and effective in labeling the informative images. We demonstrate, across several challenging benchmarks, that ALWOD significantly narrows the gap between the ODs trained on few partially labeled but strategically selected image instances and those that rely on the fully-labeled data. Our code is publicly available on https://github.com/seqam-lab/ALWOD.

****

Prototypical Kernel Learning and Open-set Foreground Perception for Generalized Few-shot Semantic Segmentation

Kai Huang, Feigege Wang, Ye Xi, Yutao Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19256-19265

Generalized Few-shot Semantic Segmentation (GFSS) extends Few-shot Semantic Segmentation (FSS) to simultaneously segment unseen classes and seen classes during evaluation. Previous works leverage additional branch or prototypical aggregation to eliminate the constrained setting of FSS. However, representation division and embedding prejudice, which heavily results in poor performance of GFSS, have not been synthetical considered. We address the aforementioned problems by jointing the prototypical kernel learning and open-set foreground perception. Specifically, a group of learnable kernels is proposed to perform segmentation with each kernel in charge of a stuff class. Then, we explore to merge the prototypical learning to the update of base-class kernels, which is consistent with the prototype knowledge aggregation of few-shot novel classes. In addition, a foreground contextual perception module cooperating with conditional bias based inference is adopted to perform class-agnostic as well as open-set foreground detection, thus to mitigate the embedding prejudice and prevent novel targets from being misclassified as background. Moreover, we also adjust our method to the Class Incremental Few-shot Semantic Segmentation (CIFSS) which takes the knowledge of novel classes in a incremental stream. Extensive experiments on PASCAL-5i and COCO-20i datasets demonstrate that our method performs better than previous state-of-the-art.

****

Simple and Effective Out-of-Distribution Detection via Cosine-based Softmax Loss

SoonCheol Noh, DongEon Jeong, Jee-Hyong Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16560-16569

Deep learning models need to detect out-of-distribution (OOD) data in the inference stage because they are trained to estimate the train distribution and infer the data sampled from the distribution. Many methods have been proposed, but they have some limitations, such as requiring additional data, input processing, or high computational cost. Moreover, most methods have hyperparameters to be set by users, which have a significant impact on the detection rate. We propose a simple and effective OOD detection method by combining the feature norm and the Mahalanobis distance obtained from classification models trained with the cosine-based softmax loss. Our method is practical because it does not use additional data for training, is about three times faster when inferencing than the methods using the input processing, and is easy to apply because it does not have any hyperparameters for OOD detection. We confirm that our method is superior to or at

least comparable to state-of-the-art OOD detection methods through the experimen
ts.
****
CFCG: Semi-Supervised Semantic Segmentation via Cross-Fusion and Contour Guidanc
e Supervision
Shuo Li, Yue He, Weiming Zhang , Wei Zhang, Xiao Tan, Junyu Han, Errui Ding, Jin
gdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2023, pp. 16348-16358
Current state-of-the-art semi-supervised semantic segmentation (SSSS) methods ty
pically adopt pseudo labeling and consistency regularization between multiple le
arners with different perturbations. Although the performance is desirable, many
 issues remain: (1) supervisions from a single learner tend to be noisy which ca
uses unreliable consistency regularization (2) existing pixel-wise confidence-sc
ore-based reliability measurement causes potential error accumulation as the tra
ining proceeds. In this paper, we propose a novel SSSS framework, called CFCG, w
hich combines cross-fusion and contour guidance supervision to tackle these issu
es.
 Concretely, we adopt both image-level and feature-level perturbations to expand
 feature distribution thus pushing the potential limits of consistency regulariz
ation. Then, two particular modules are proposed to enable effective semi-superv
ised learning under heavy coherent perturbations. Firstly, Cross-Fusion Supervis
ion (CFS) mechanism leverages multiple learners to enhance the quality of pseudo
 labels. Secondly, we introduce an adaptive contour guidance module (ACGM) to ef
fectively identify unreliable spatial regions in pseudo labels. Finally, our pro
posed CFCG achieves gains of mIoU +1.40%, +0.89% with a single learner and +1.85
%, +1.33% by fusion inference on PASCAL VOC 2012 and on Cityscapes respectively
under 1/8 protocols, clearly surpassing previous methods and reaching the state-
of-the-art.
****
CHAMPAGNE: Learning Real-world Conversation from Large-Scale Web Videos
Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, Youngjae Yu; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15498-
15509
Visual information is central to conversation: body gestures and physical behavi
our, for example, contribute to meaning that transcends words alone. To date, ho
wever, most neural conversational models are limited to just text. We introduce
CHAMPAGNE, a generative model of conversations that can account for visual conte
xts. To train CHAMPAGNE, we collect and release YTD-18M, a large-scale corpus of
 18M video-based dialogues. YTD-18M is constructed from web videos: crucial to o
ur data collection pipeline is a pretrained language model that converts error-p
rone automatic transcripts to a cleaner dialogue format while maintaining meanin
g.
 Human evaluation reveals that YTD-18M is more sensible and specific than prior
resources (MMDialog, 1M dialogues), while maintaining visual-groundedness. Exper
iments demonstrate that 1) CHAMPAGNE learns to conduct conversation from YTD-18M
; and 2) when fine-tuned, it achieves state-of-the-art results on four vision-la
nguage tasks focused on real-world conversations. We release data, models, and c
ode.
****
SLAN: Self-Locator Aided Network for Vision-Language Understanding
Jiang-Tian Zhai, Qi Zhang, Tong Wu, Xing-Yu Chen, Jiang-Jiang Liu, Ming-Ming Che
ng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC
V), 2023, pp. 21949-21958
Learning fine-grained interplay between vision and language contributes to a mor
e accurate understanding for Vision-Language tasks. However, it remains challeng
ing to extract key image regions according to the texts for semantic
 alignments. Most existing works are either limited by text-agnostic and redunda
nt regions obtained with the frozen detectors, or failing to scale further due t
o their heavy reliance on scarce grounding (gold) data to pre-train detectors. T
o

solve these problems, we propose Self-Locator Aided Network (SLAN) for vision-language understanding tasks without any extra gold data. SLAN consists of a region filter and a region adaptor to localize regions of interest conditioned on different texts. By aggregating vision-language information, the region filter selects key regions and the region adaptor updates their coordinates with text guidance. With detailed region-word alignments, SLAN can be easily generalized to many downstream tasks. It achieves fairly competitive results on five vision-language understanding tasks (e.g., 85.7% and 69.2% on COCO image-to-text and text-to-image retrieval, surpassing previous SOTA methods). SLAN also demonstrates strong zero-shot and fine-tuned transferability to two localization tasks.
****
S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces
Haoyu Wu, Alexandros Graikos, Dimitris Samaras; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3556-3568
Neural rendering of implicit surfaces performs well in 3D vision applications. However, it requires dense input views as supervision. When only sparse input images are available, output quality drops significantly due to the shape-radiance ambiguity problem. We note that this ambiguity can be constrained when a 3D point is visible in multiple views, as is the case in multi-view stereo (MVS). We thus propose to regularize neural rendering optimization with an MVS solution. The use of an MVS probability volume and a generalized cross entropy loss leads to a noise-tolerant optimization process. In addition, neural rendering provides global consistency constraints that guide the MVS depth hypothesis sampling and thus improves MVS performance. Given only three sparse input views, experiments show that our method not only outperforms generic neural rendering models by a large margin but also significantly increases the reconstruction quality of MVS models.
****
Anomaly Detection using Score-based Perturbation Resilience
Woosang Shin, Jonghyeon Lee, Taehan Lee, Sangmoon Lee, Jong Pil Yun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23372-23382
Unsupervised anomaly detection is widely studied for industrial applications since it is difficult to obtain anomalous data. In particular, reconstruction-based anomaly detection can be a feasible solution if there is no option to use external knowledge, such as extra datasets or pre-trained models. However, reconstruction-based methods have limited utility due to poor detection performance. A score-based model, also known as a denoising diffusion model, recently has shown a high sample quality in the generation task. In this paper, we propose a novel unsupervised anomaly detection method leveraging the score-based model. This method promises good performance without external knowledge. The score, a gradient of the log-likelihood, has a property that is available for anomaly detection. The samples on the data manifold can be restored instantly by the score, even if they are randomly perturbed. We call this a score-based perturbation resilience. On the other hand, the samples that deviate from the manifold cannot be restored in the same way. The variation of resilience depending on the sample position can be an indicator to discriminate anomalies. We derive this statement from a geometric perspective. Our method shows superior performance on three benchmark datasets for industrial anomaly detection. Specifically, on MVTec AD, we achieve image-level AUROC of 97.7% and pixel-level AUROC of 97.4% outperforming previous works that do not use external knowledge.
****
Generating Visual Scenes from Touch
Fengyu Yang, Jiacheng Zhang, Andrew Owens; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22070-22080
An emerging line of work has sought to generate plausible imagery from touch. Existing approaches, however, tackle only narrow aspects of the visuo-tactile synthesis problem, and lag significantly behind the quality of cross-modal synthesis methods in other domains. We draw on recent advances in latent diffusion to create a model for synthesizing images from tactile signals (and vice versa) and ap

ply it to a number of visuo-tactile synthesis tasks. Using this model, we signif icantly outperform prior work on the tactile-driven stylization problem, i.e., m anipulating an image to match a touch signal, and we are the first to successful ly generate images from touch without additional sources of information about th e scene. We also successfully use our model to address two novel synthesis probl ems: generating images that do not contain the touch sensor or the hand holding it, and estimating an image's shading from its reflectance and touch.
****
DeformToon3D: Deformable Neural Radiance Fields for 3D Toonification
Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Zi wei Liu, Chen Change Loy; Proceedings of the IEEE/CVF International Conference o n Computer Vision (ICCV), 2023, pp. 9144-9154
In this paper, we address the challenging problem of 3D toonification, which inv olves transferring the style of an artistic domain onto a target 3D face with st ylized geometry and texture. Although fine-tuning a pre-trained 3D GAN on the ar tistic domain can produce reasonable performance, this strategy has limitations in the 3D domain. In particular, fine-tuning can deteriorate the original GAN la tent space, which affects subsequent semantic editing, and requires independent optimization and storage for each new style, limiting flexibility and efficient deployment. To overcome these challenges, we propose DeformToon3D, an effective toonification framework tailored for hierarchical 3D GAN. Our approach decompose s 3D toonification into subproblems of geometry and texture stylization to bette r preserve the original latent space. Specifically, we devise a novel StyleField that predicts conditional 3D deformation to align a real-space NeRF to the styl e space for geometry stylization. Thanks to the StyleField formulation, which al ready handles geometry stylization well, texture stylization can be achieved con veniently via adaptive style mixing that injects information of the artistic dom ain into the decoder of the pre-trained 3D GAN. Due to the unique design, our me thod enables flexible style degree control and shape-texture-specific style swap . Furthermore, we achieve efficient training without any real-world 2D-3D traini ng pairs but proxy samples synthesized from off-the-shelf 2D toonification model s.
****
SatlasPretrain: A Large-Scale Dataset for Remote Sensing Image Understanding
Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, Aniruddha Kembhavi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16772-16782
Remote sensing images are useful for a wide variety of planet monitoring applica tions, from tracking deforestation to tackling illegal fishing. The Earth is ext remely diverse---the amount of potential tasks in remote sensing images is massi ve, and the sizes of features range from several kilometers to just tens of cent imeters. However, creating generalizable computer vision methods is a challenge in part due to the lack of a large-scale dataset that captures these diverse fea tures for many tasks. In this paper, we present SatlasPretrain, a remote sensing dataset that is large in both breadth and scale, combining Sentinel-2 and NAIP images with 302M labels under 137 categories and seven label types. We evaluate eight baselines and a proposed method on SatlasPretrain, and find that there is substantial room for improvement in addressing research challenges specific to r emote sensing, including processing image time series that consist of images fro m very different types of sensors, and taking advantage of long-range spatial co ntext. Moreover, we find that pre-training on SatlasPretrain substantially impro ves performance on downstream tasks, increasing average accuracy by 18% over Ima geNet and 6% over the next best baseline. The dataset, pre-trained model weights , and code are available at https://satlas-pretrain.allen.ai/.
****
Empowering Low-Light Image Enhancer through Customized Learnable Priors
Naishan Zheng, Man Zhou, Yanmeng Dong, Xiangyu Rui, Jie Huang, Chongyi Li, Feng Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (I CCV), 2023, pp. 12559-12569
Deep neural networks have achieved remarkable progress in enhancing low-light im

ages by improving their brightness and eliminating noise. However, most existing
 methods construct end-to-end mapping networks heuristically, neglecting the int
rinsic prior of image enhancement task and lacking transparency and interpretabi
lity. Although some unfolding solutions have been proposed to relieve these issu
es, they rely on proximal operator networks that deliver ambiguous and implicit
priors. In this work, we propose a paradigm for low-light image enhancement that
 explores the potential of customized learnable priors to improve the transparen
cy of the deep unfolding paradigm.Motivated by the powerful feature representati
on capability of Masked Autoencoder (MAE), we customize MAE-based illumination a
nd noise priors and redevelop them from two perspectives: 1) structure flow: we
train the MAE from a normal-light image to its illumination properties and then
embed it into the proximal operator design of the unfolding architecture; and
 2) optimization flow: we train MAE from a normal-light image to its gradient re
presentation and then employ it as a regularization term to constrain noise in t
he model output. These designs improve the interpretability and representation c
apability of the model. Extensive experiments on multiple low-light image enhanc
ement datasets demonstrate the superiority of our proposed paradigm over state-o
f-the-art methods. Code is available at https://github.com/zheng980629/CUE.
****
TextManiA: Enriching Visual Feature by Text-driven Manifold Augmentation
Moon Ye-Bin, Jisoo Kim, Hongyeob Kim, Kilho Son, Tae-Hyun Oh; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2526-253
7
We propose TextManiA, a text-driven manifold augmentation method that semantical
ly enriches visual feature spaces, regardless of class distribution. TextManiA a
ugments visual data with intra-class semantic perturbation by exploiting easy-to
-understand visually mimetic words, i.e., attributes. This work is built on an i
nteresting hypothesis that general language models, e.g., BERT and GPT, encompas
s visual information to some extent, even without training on visual training da
ta. Given the hypothesis, TextManiA transfers pre-trained text representation ob
tained from a well-established large language encoder to a target visual feature
 space being learned. Our extensive analysis hints that the language encoder ind
eed encompasses visual information at least useful to augment visual representat
ion. Our experiments demonstrate that TextManiA is particularly powerful in scar
ce samples with class imbalance as well as even distribution. We also show compa
tibility with the label mix-based approaches in evenly distributed scarce data.
****
Guiding Image Captioning Models Toward More Specific Captions
Simon Kornblith, Lala Li, Zirui Wang, Thao Nguyen; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 15259-15269
Image captioning is conventionally formulated as the task of generating captions
 that match the conditional distribution of reference image-caption pairs. Howev
er, reference captions in standard captioning datasets are short and may not uni
quely identify the images they describe. These problems are further exacerbated
when models are trained directly on image-alt text pairs collected from the inte
rnet. In this work, we show that it is possible to generate more specific captio
ns with minimal changes to the training process. We implement classifier-free gu
idance (Ho & Salimans, 2021) for an autoregressive captioning model by fine-tuni
ng it to estimate both conditional and unconditional distributions over captions
. The guidance scale applied at decoding controls a trade-off between maximizing
 p(caption|image) and p(caption|image). Compared to standard greedy decoding, de
coding with a guidance scale of 2 substantially improves reference-free metrics
such as CLIPScore (0.808 vs. 0.775) and caption->image retrieval performance in
the CLIP embedding space (recall@1 44.6% vs. 26.5%), but worsens standard refere
nce-based captioning metrics (e.g., CIDEr 78.6 vs 126.1). We further explore the
 use of language models to guide the decoding process, obtaining small improveme
nts over the Pareto frontier of reference-free vs. reference-based captioning me
trics that arises from classifier-free guidance, and substantially improving the
 grammaticality of captions generated from a model trained only on minimally cur
ated web data.

****
Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images
Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, Roy Schwartz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2616-2627
Weird, unusual, and uncanny images pique the curiosity of observers because they challenge commonsense. For example, an image released during the 2022 world cup depicts the famous soccer stars Lionel Messi and Cristiano Ronaldo playing chess, which playfully violates our expectation that their competition should occur on the football field.
Humans can easily recognize and interpret these unconventional images, but can AI models do the same?
We introduce WHOOPS!, a new dataset and benchmark for visual commonsense. The dataset is comprised of purposefully commonsense-defying images created by designers using publicly-available image generation tools like Midjourney.
We consider several tasks posed over the dataset.
In addition to image captioning, cross-modal matching, and visual question answering, we introduce a difficult explanation generation task, where models must identify and explain why a given image is unusual. Our results show that state-of-the-art models such as GPT3 and BLIP2 still lag behind human performance on WHOOPS!.
We hope our dataset will inspire the development of AI models with stronger visual commonsense reasoning abilities.
****
Consistent Depth Prediction for Transparent Object Reconstruction from RGB-D Camera
Yuxiang Cai, Yifan Zhu, Haiwei Zhang, Bo Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3459-3468
Transparent objects are commonly seen in indoor scenes but are hard to estimate. Currently, commercial depth cameras face difficulties in estimating the depth of transparent objects due to the light reflection and refraction on their surface. As a result, they tend to make a noisy and incorrect depth value for transparent objects. These incorrect depth data make the traditional RGB-D SLAM method fails in reconstructing the scenes that contain transparent objects. An exact depth value of the transparent object is required to restore in advance and it is essential that the depth value of the transparent object must keep consistent in different views, or the reconstruction result will be distorted. Previous depth prediction methods of transparent objects can restore these missing depth values but none of them can provide a good result in reconstruction due to the inconsistency prediction. In this work, we propose a real-time reconstruction method using a novel stereo-based depth prediction network to keep the consistency of depth prediction in a sequence of images. Because there is no video dataset about transparent objects currently to train our model, we construct a synthetic RGB-D video dataset with different transparent objects. Moreover, to test generalization capability, we capture video from real scenes using the RealSense D435i RGB-D camera. We compare the metrics on our dataset and SLAM reconstruction results in both synthetic scenes and real scenes with the previous methods. Experiments show our significant improvement in accuracy on depth prediction and scene reconstruction.
****
DReg-NeRF: Deep Registration for Neural Radiance Fields
Yu Chen, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22703-22713
Although Neural Radiance Fields (NeRF) is popular in the computer vision community recently, registering multiple NeRFs has yet to gain much attention. Unlike the existing work, NeRF2NeRF, which is based on traditional optimization methods and needs human annotated keypoints, we propose DReg-NeRF to solve the NeRF registration problem on object-centric scenes without human intervention. After training NeRF models, our DReg-NeRF first extracts features from the occupancy grid

in NeRF. Subsequently, our DReg-NeRF utilizes a transformer architecture with se
lf-attention and cross-attention layers to learn the relations between pairwise
NeRF blocks. In contrast to state-of-the-art (SOTA) point cloud registration met
hods, the decoupled correspondences are supervised by surface fields without any
 ground truth overlapping labels. We construct a novel view synthesis dataset wi
th 1,700+ 3D objects obtained from Objaverse to train our network. When evaluate
d on the test set, our proposed method beats the SOTA point cloud registration m
ethods by a large margin with a mean RPE = 9.67* and a mean RTE = 0.038. Our cod
e is available at https://github.com/AIBluefisher/DReg-NeRF.
****
DETR Does Not Need Multi-Scale or Locality Design
Yutong Lin, Yuhui Yuan, Zheng Zhang, Chen Li, Nanning Zheng, Han Hu; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6
545-6554
This paper presents an improved DETR detector that maintains a "plain" nature: u
sing a single-scale feature map and global cross-attention calculations without
specific locality constraints, in contrast to previous leading DETR-based detect
ors that reintroduce architectural inductive biases of multi-scale and locality
into the decoder. We show that two simple technologies are surprisingly effectiv
e within a plain design to compensate for the lack of multi-scale feature maps a
nd locality constraints. The first is a box-to-pixel relative position bias (Box
RPB) term added to the cross-attention formulation, which well guides each query
 to attend to the corresponding object region while also providing encoding flex
ibility. The second is masked image modeling (MIM)-based backbone pre-training w
hich helps learn representation with fine-grained localization ability and prove
s crucial for remedying dependencies on the multi-scale feature maps.
 By incorporating these technologies and recent advancements in training and pro
blem formation, the improved "plain" DETR showed exceptional improvements over t
he original DETR detector. By leveraging the Object365 dataset for pre-training,
 it achieved 63.9 mAP accuracy using a Swin-L backbone, which is highly competit
ive with state-of-the-art detectors which all heavily rely on multi-scale featur
e maps and region-based feature extraction. Code will be available at https://gi
thub.com/impiga/Plain-DETR.
****
Towards Effective Instance Discrimination Contrastive Loss for Unsupervised Doma
in Adaptation
Yixin Zhang, Zilei Wang, Junjie Li, Jiafan Zhuang, Zihan Lin; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11388-11
399
Domain adaptation (DA) aims to transfer knowledge from a label-rich source domai
n to a related but label-scarce target domain. Recently, increasing research has
 focused on exploring data structure of the target domain. In light of the recen
t success of Instance Discrimination Contrastive (IDCo) loss in self-supervised
learning, we try directly applying it to domain adaptation tasks. However, the i
mprovement is very limited, which motivates us to rethink its underlying limitat
ions for domain adaptation tasks. An intuitive limitation is that a pair of samp
les belonging to the same class could be treated as negatives. Here we argue tha
t using low-confidence samples to construct positive and negative pairs can alle
viate this issue and is more suitable for IDCo loss. Another limitation is that
IDCo loss cannot capture enough semantic information. We address this by introdu
cing domain-invariant and accurate semantic information from classifier weights
and input data. Specifically, we propose a class relationship enhanced features.
 It uses probability weighted class prototpyes as the input features of IDCo los
s, which can implicitly transfer the domain-invariant class relationship. We fur
ther propose a target-dominated cross-domain mixup that can incorporate accurate
 semantic information from the source domain. We evaluate the proposed method in
 unsupervised DA and other DA settings, and extensive experimental results revea
l that our method can make IDCo loss more effective and achieve state-of-the-art
 performance.
****

ClusT3: Information Invariant Test-Time Training
Gustavo A. Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Ismail Ben Ayed, Christian Desrosiers; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6136-6145
Deep Learning models have shown remarkable performance in a broad range of vision tasks. However, they are often vulnerable against domain shifts at test-time. Test-time training (TTT) methods have been developed in an attempt to mitigate these vulnerabilities, where a secondary task is solved at training time simultaneously with the main task, to be later used as an self-supervised proxy task at test-time. In this work, we propose a novel unsupervised TTT technique based on the maximization of Mutual Information between multi-scale feature maps and a discrete latent representation, which can be integrated to the standard training as an auxiliary clustering task. Experimental results demonstrate competitive classification performance on different popular test-time adaptation benchmarks. The code can be found at: https://github.com/dosowiechi/ClusT3.git
****
FrozenRecon: Pose-free 3D Scene Reconstruction with Frozen Depth Models
Guangkai Xu, Wei Yin, Hao Chen, Chunhua Shen, Kai Cheng, Feng Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9310-9320
3D scene reconstruction is a long-standing vision task. Existing approaches can be categorized into geometry-based and learning-based methods. The former leverages multi-view geometry but may face catastrophic failures due to the reliance on accurate pixel correspondence across views, while the latter mitigates these issues by learning 2D or 3D representation directly. However, without a large-scale video or 3D training data, it can hardly be generalized to diverse real-world scenarios due to the presence of tens of millions or even billions of optimization parameters in the deep network.
Recently, robust monocular depth estimation models trained with large-scale datasets have been proven to possess weak 3D geometry prior, but they are insufficient for reconstruction due to the unknown camera parameters, the affine-invariant property, and inter-frame inconsistency. To address these issues, we propose a novel test-time optimization approach that can transfer the robustness of affine-invariant depth models such as LeReS to challenging diverse scenes while ensuring inter-frame consistency, with only dozens of parameters to optimize per video frame. Specifically, our approach involves freezing the pre-trained affine-invariant depth model's depth predictions, rectifying them by optimizing the unknown scale-shift values with a geometric consistency alignment module, and employing the resulting scale-consistent depth maps to robustly obtain camera poses and achieve dense scene reconstruction, even in low-texture regions. Experiments show that our method achieves state-of-the-art cross-dataset reconstruction on five zero-shot testing datasets. Code is available at: https://aim-uofa.github.io/FrozenRecon/
****
Affective Image Filter: Reflecting Emotions from Text to Images
Shuchen Weng, Peixuan Zhang, Zheng Chang, Xinlong Wang, Si Li, Boxin Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10810-10819
Understanding the emotions in text and presenting them visually is a very challenging problem that requires a deep understanding of natural language and high-quality image synthesis simultaneously. In this work, we propose Affective Image Filter (AIF), a novel model that is able to understand the visually-abstract emotions from the text and reflect them to visually-concrete images with appropriate colors and textures. We build our model based on the multi-modal transformer architecture, which unifies both images and texts into tokens and encodes the emotional prior knowledge. Various loss functions are proposed to understand complex emotions and produce appropriate visualization. In addition, we collect and contribute a new dataset with abundant aesthetic images and emotional texts for training and evaluating the AIF model. We carefully design four quantitative metrics and conduct a user study to comprehensively evaluate the performance, which de

monstrates our AIF model outperforms state-of-the-art methods and could evoke specific emotional responses from human observers.
****

Content-Aware Local GAN for Photo-Realistic Super-Resolution
JoonKyu Park, Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10585-10594
Recently, GAN has successfully contributed to making single-image super-resolution (SISR) methods produce more realistic images. However, natural images have complex distribution in the real world, and a single classifier in the discriminator may not have enough capacity to classify real and fake samples, making the preceding SR network generate unpleasing noise and artifacts. To solve the problem, we propose a novel content-aware local GAN framework, CAL-GAN, which processes a large and complicated distribution of real-world images by dividing them into smaller subsets based on similar contents. Our mixture of classifiers (MoC) design allocates different super-resolved patches to corresponding expert classifiers. Additionally, we introduce novel routing and orthogonality loss terms so that different classifiers can handle various contents and learn separable features. By feeding similar distributions into the corresponding specialized classifiers, CAL-GAN enhances the representation power of existing super-resolution models, achieving state-of-the-art perceptual performance on standard benchmarks and real-world images without modifying the generator-side architecture.
****

Structure-Aware Surface Reconstruction via Primitive Assembly
Jingen Jiang, Mingyang Zhao, Shiqing Xin, Yanchao Yang, Hanxiao Wang, Xiaohong Jia, Dong-Ming Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14171-14180
We propose a novel and efficient method for reconstructing manifold surfaces from point clouds. Unlike previous approaches that use dense implicit reconstructions or piecewise approximations and overlook inherent structures like quadrics in CAD models, our method faithfully preserves these quadric structures by assembling primitives. To achieve high-quality primitive extraction, we use a variational shape approximation, followed by a mesh arrangement for space partitioning and candidate primitive patches generation. We then introduce an effective pruning mechanism to classify candidate primitive patches as active or inactive, and further prune inactive patches to reduce the search space and speed up surface extraction significantly. Finally, the optimal active patches are computed by a binary linear programming and assembled as manifold and watertight surfaces. We perform extensive experiments on a wide range of CAD objects to validate its effectiveness.
****

FineDance: A Fine-grained Choreography Dataset for 3D Full Body Dance Generation
Ronghui Li, Junfan Zhao, Yachao Zhang, Mingyang Su, Zeping Ren, Han Zhang, Yansong Tang, Xiu Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10234-10243
Generating full-body and multi-genre dance sequences from given music is a challenging task, due to the limitations of existing datasets and the inherent complexity of the fine-grained hand motion and dance genres. To address these problems, we propose FineDance, which contains 14.6 hours of music-dance paired data, with fine-grained hand motions, fine-grained genres (22 dance genres), and accurate posture. To the best of our knowledge, FineDance is the largest music-dance paired dataset with the most dance genres. Additionally, to address monotonous and unnatural hand movements existing in previous methods, we propose a full-body dance generation network, which utilizes the diverse generation capabilities of the diffusion model to solve monotonous problems, and use expert nets to solve unreal problems. To further enhance the genrematching and long-term stability of generated dances, we propose a Genre&Coherent aware Retrieval Module. Besides, we propose a new metric named Genre Matching Score to measure the genre matching between dance and music. Quantitative and qualitative experiments demonstrate the quality of FineDance, and the state-of-the-art performance of FineNet.

****
AssetField: Assets Mining and Reconfiguration in Ground Feature Plane Representation
Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Bo Dai, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3251-3261
Both indoor and outdoor environments are inherently structured and repetitive. Traditional modeling pipelines keep an asset library storing unique object templates, which is both versatile and memory efficient in practice. Inspired by this observation, we propose AssetField, a novel neural scene representation that learns a set of object-aware ground feature planes to represent the scene, where an asset library storing template feature patches can be constructed in an unsupervised manner. Unlike existing methods which require object masks to query spatial points for object editing, our ground feature plane representation offers a natural visualization of the scene in the bird-eye view, allowing a variety of operations (e.g. translation, duplication, deformation) on objects to configure a new scene. With the template feature patches, group editing is enabled for scenes with many recurring items to avoid repetitive work on object individuals. We show that AssetField not only achieves competitive performance for novel-view synthesis but also generates realistic renderings for new scene configurations.
****
Improving Online Lane Graph Extraction by Object-Lane Clustering
Yigit Baran Can, Alexander Liniger, Danda Pani Paudel, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8591-8601
Autonomous driving requires accurate local scene understanding information. To this end, autonomous agents deploy object detection and online BEV lane graph extraction methods as a part of their perception stack. In this work, we propose an architecture and loss formulation to improve the accuracy of local lane graph estimates by using 3D object detection outputs. The proposed method learns to assign the objects to centerlines by considering the centerlines as cluster centers and the objects as data points to be assigned a probability distribution over the cluster centers. This training scheme ensures direct supervision on the relationship between lanes and objects, thus leading to better performance. The proposed method improves lane graph estimation substantially over state-of-the-art methods. The extensive ablations show that our method can achieve significant performance improvements by using the outputs of existing 3D object detection methods. Since our method uses the detection outputs rather than detection method intermediate representations, a single model of our method can use any detection method at test time. The code will be made publicly available.
****
SAGA: Spectral Adversarial Geometric Attack on 3D Meshes
Tomer Stolik, Itai Lang, Shai Avidan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4284-4294
A triangular mesh is one of the most popular 3D data representations. As such, the deployment of deep neural networks for mesh processing is widely spread and is increasingly attracting more attention. However, neural networks are prone to adversarial attacks, where carefully crafted inputs impair the model's functionality. The need to explore these vulnerabilities is a fundamental factor in the future development of 3D-based applications. Recently, mesh attacks were studied on the semantic level, where classifiers are misled to produce wrong predictions. Nevertheless, mesh surfaces possess complex geometric attributes beyond their semantic meaning, and their analysis often includes the need to encode and reconstruct the geometry of the shape. We propose a novel framework for a geometric adversarial attack on a 3D mesh autoencoder. In this setting, an adversarial input mesh deceives the autoencoder by forcing it to reconstruct a different geometric shape at its output. The malicious input is produced by perturbing a clean shape in the spectral domain. Our method leverages the spectral decomposition of the mesh along with additional mesh-related properties to obtain visually credible results that consider the delicacy of surface distortions.

****
All in Tokens: Unifying Output Space of Visual Tasks via Soft Token
Jia Ning, Chen Li, Zheng Zhang, Chunyu Wang, Zigang Geng, Qi Dai, Kun He, Han Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19900-19910
We introduce AiT, a unified output representation for various vision tasks, which is a crucial step towards general-purpose vision task solvers. Despite the challenges posed by the high-dimensional and task-specific outputs, we showcase the potential of using discrete representation (VQ-VAE) to model the dense outputs of many computer vision tasks as a sequence of discrete tokens. This is inspired by the established ability of VQ-VAE to conserve the structures spanning multiple pixels using few discrete codes. To that end, we present a modified shallower architecture for VQ-VAE that improves efficiency while keeping prediction accuracy. Our approach also incorporates uncertainty into the decoding process by using a soft fusion of the codebook entries, providing a more stable training process, which notably improved prediction accuracy. Our evaluation of AiT on depth estimation and instance segmentation tasks, with both continuous and discrete labels, demonstrates its superiority compared to other unified models. The code and models are available at https://github.com/SwinTransformer/AiT.
****
Learning Navigational Visual Representations with Semantic Map Supervision
Yicong Hong, Yang Zhou, Ruiyi Zhang, Franck Dernoncourt, Trung Bui, Stephen Gould, Hao Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3055-3067
Being able to perceive the semantics and the spatial structure of the environment is essential for visual navigation of a household robot. However, most existing works only employ visual backbones pre-trained either with independent images for classification or with self-supervised learning methods to adapt to the indoor navigation domain, both neglecting the spatial relationships that are essential to the learning of navigation. Inspired by the behavior that human naturally build semantically and spatially meaningful cognitive maps in their brain during navigation, in this paper, we propose a novel navigational-specific visual representation learning method by contrasting the agent's egocentric views and semantic maps (Ego^2-Map). We apply the visual transformer as the backbone encoder and train the model with data collected from the large-scale Habitat-Matterport3D environments. Ego^2-Map learning transfers the compact and rich information from a map, such as objects, structure and transition, to the agent's egocentric representations for navigation. Experiments show that agents using our learned representations on object-goal navigation outperforms recent visual pre-training methods. Moreover, our representations lead to a significant improvement in vision-and-language navigation in continuous environments for both high-level and low-level action spaces, achieving new state-of-the-art results of 47% SR and 41% SPL on the test server.
****
LDL: Line Distance Functions for Panoramic Localization
Junho Kim, Changwoon Choi, Hojun Jang, Young Min Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17882-17892
We introduce LDL, a fast and robust algorithm that localizes a panorama to a 3D map using line segments. LDL focuses on the sparse structural information of lines in the scene, which is robust to illumination changes and can potentially enable efficient computation. While previous line-based localization approaches tend to sacrifice accuracy or computation time, our method effectively observes the holistic distribution of lines within panoramic images and 3D maps. Specifically, LDL matches the distribution of lines with 2D and 3D line distance functions, which are further decomposed along principal directions of lines to increase the expressiveness. The distance functions provide coarse pose estimates by comparing the distributional information, where the poses are further optimized using conventional local feature matching. As our pipeline solely leverages line geometry and local features, it does not require costly additional training of line-specific features or correspondence matching. Nevertheless, our method demonstrat

es robust performance on challenging scenarios including object layout changes, illumination shifts, and large-scale scenes, while exhibiting fast pose search terminating within a matter of milliseconds. We thus expect our method to serve as a practical solution for line-based localization, and complement the well-established point-based paradigm.
****

TransTIC: Transferring Transformer-based Image Compression from Human Perception to Machine Perception

Yi-Hsin Chen, Ying-Chieh Weng, Chia-Hao Kao, Cheng Chien, Wei-Chen Chiu, Wen-Hsiao Peng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23297-23307

This work aims for transferring a Transformer-based image compression codec from human perception to machine perception without fine-tuning the codec. We propose a transferable Transformer-based image compression framework, termed TransTIC. Inspired by visual prompt tuning, TransTIC adopts an instance-specific prompt generator to inject instance-specific prompts to the encoder and task-specific prompts to the decoder. Extensive experiments show that our proposed method is capable of transferring the base codec to various machine tasks and outperforms the competing methods significantly. To our best knowledge, this work is the first attempt to utilize prompting on the low-level image compression task.
****

CHORUS : Learning Canonicalized 3D Human-Object Spatial Relations from Unbounded Synthesized Images

Sookwan Han, Hanbyul Joo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15835-15846

We present a method for teaching machines to understand and model the underlying spatial common sense of diverse human-object interactions in 3D in a self-supervised way. This is a challenging task, as there exist specific manifolds of the interactions that can be considered human-like and natural, but the human pose and the geometry of objects can vary even for similar interactions. Such diversity makes the annotating task of 3D interactions difficult and hard to scale, which limits the potential to reason about that in a supervised way. One way of learning the 3D spatial relationship between humans and objects during interaction is by showing multiple 2D images captured from different viewpoints when humans interact with the same type of objects. The core idea of our method is to leverage a generative model that produces high-quality 2D images from an arbitrary text prompt input as an "unbounded" data generator with effective controllability and view diversity. Despite its imperfection of the image quality over real images, we demonstrate that the synthesized images are sufficient to learn the 3D human-object spatial relations. We present multiple strategies to leverage the synthesized images, including (1) the first method to leverage a generative image model for 3D human-object spatial relation learning; (2) a framework to reason about the 3D spatial relations from inconsistent 2D cues in a self-supervised manner via 3D occupancy reasoning with pose canonicalization; (3) semantic clustering to disambiguate different types of interactions with the same object types; and (4) a novel metric to assess the quality of 3D spatial learning of interaction. Project Page: https://jellyheadandrew.github.io/projects/chorus
****

Shortcut-V2V: Compression Framework for Video-to-Video Translation Based on Temporal Redundancy Reduction

Chaeyeon Chung, Yeojeong Park, Seunghwan Choi, Munkhsoyol Ganbat, Jaegul Choo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7612-7622

Video-to-video translation aims to generate video frames of a target domain from an input video. Despite its usefulness, the existing networks require enormous computations, necessitating their model compression for wide use. While there exist compression methods that improve computational efficiency in various image/video tasks, a generally applicable compression method for video-to-video translation has not been studied much. In response, we present Shortcut-V2V, a general-purpose compression framework for video-to-video translation. Shortcut-V2V avoid

s full inference for every neighboring video frame by approximating the intermediate features of a current frame from those of the previous frame. Moreover, in our framework, a newly-proposed block called AdaBD adaptively blends and deforms features of neighboring frames, which makes more accurate predictions of the intermediate features possible. We conduct quantitative and qualitative evaluations using well-known video-to-video translation models on various tasks to demonstrate the general applicability of our framework. The results show that Shortcut-V2V achieves comparable performance compared to the original video-to-video translation model while saving 3.2-5.7x computational cost and 7.8-44x memory at test time. Our code and videos are available at https://shortcut-v2v.github.io/.
****
ViLLA: Fine-Grained Vision-Language Representation Learning from Real-World Data
Maya Varma, Jean-Benoit Delbrouck, Sarah Hooper, Akshay Chaudhari, Curtis Langlotz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22225-22235
Vision-language models (VLMs), such as CLIP and ALIGN, are generally trained on datasets consisting of image-caption pairs obtained from the web. However, real-world multimodal datasets, such as healthcare data, are significantly more complex: each image (e.g. X-ray) is often paired with text (e.g. physician report) that describes many distinct attributes occurring in fine-grained regions of the image. We refer to these samples as exhibiting high pairwise complexity, since each image-text pair can be decomposed into a large number of region-attribute pairings. The extent to which VLMs can capture fine-grained relationships between image regions and textual attributes when trained on such data has not been previously evaluated. The first key contribution of this work is to demonstrate through systematic evaluations that as the pairwise complexity of the training dataset increases, standard VLMs struggle to learn region-attribute relationships, exhibiting performance degradations of up to 37% on retrieval tasks. In order to address this issue, we introduce ViLLA as our second key contribution. ViLLA, which is trained to capture fine-grained region-attribute relationships from complex datasets, involves two components: (a) a lightweight, self-supervised mapping model to decompose image-text samples into region-attribute pairs, and (b) a contrastive VLM to learn representations from generated region-attribute pairs. We demonstrate with experiments across four domains (synthetic, product, medical, and natural images) that ViLLA outperforms comparable VLMs on fine-grained reasoning tasks, such as zero-shot object detection (up to 3.6 AP50 points on COCO and 0.6 mAP points on LVIS) and retrieval (up to 14.2 R-Precision points).
****
SG-Former: Self-guided Transformer with Evolving Token Reallocation
Sucheng Ren, Xingyi Yang, Songhua Liu, Xinchao Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6003-6014
Vision Transformer has demonstrated impressive success across various vision tasks. However, its heavy computation cost, which grows quadratically with respect to the token sequence length, largely limits its power in handling large feature maps. To alleviate the computation cost, previous works rely on either fine-grained self-attentions restricted to local small regions, or global self-attentions but to shorten the sequence length resulting in coarse granularity. In this paper, we propose a novel model, termed as Self-guided Transformer (SG-Former), towards effective global self-attention with adaptive fine granularity. At the heart of our approach is to utilize a significance map, which is estimated through hybrid-scale self-attention and evolves itself during training, to reallocate tokens based on the significance of each region. Intuitively, we assign more tokens to the salient regions for achieving fine-grained attention, while allocating fewer tokens to the minor regions in exchange for efficiency and global receptive fields. The proposed SG-Former achieves performance superior to state of the art: our base size model achieves 84.7% Top-1 accuracy on ImageNet-1K, 51.2mAP bbAP on CoCo, 52.7mIoU on ADE20K surpassing the Swin Transformer by +1.3% / +2.7 mAP/ +3 mIoU, with lower computation costs and fewer parameters.
****
Towards Unifying Medical Vision-and-Language Pre-Training via Soft Prompts

Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, Xiang Wan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23403-23413
Medical vision-and-language pre-training (Med-VLP) has shown promising improvements on many downstream medical tasks owing to its applicability to extracting generic representations from medical images and texts. Practically, there exist two typical types, i.e., the fusion-encoder type and the dual-encoder type, depending on whether a heavy fusion module is used. The former is superior at multi-modal tasks owing to the sufficient interaction between modalities; the latter is good at uni-modal and cross-modal tasks due to the single-modality encoding ability. To take advantage of these two types, we propose an effective yet straightforward scheme named PTUnifier to unify the two types. We first unify the input format by introducing visual and textual prompts, which serve as DETR-like queries that assist in extracting features when one of the modalities is missing. By doing so, a single model could serve as a foundation model that processes various tasks adopting different input formats (i.e., image-only, text-only, and image-text-pair). Furthermore, we construct a prompt pool (instead of static ones) to improve diversity and scalability, enabling queries conditioned on different input instances. Experimental results show that our approach achieves state-of-the-art results on a broad range of tasks, spanning uni-modal tasks (i.e., image/text classification and text summarization), cross-modal tasks (i.e., image-to-text generation and image-text/text-image retrieval), and multi-modal tasks (i.e., visual question answering), demonstrating the effectiveness of our approach. Note that the adoption of prompts is orthogonal to most existing Med-VLP approaches and could be a beneficial and complementary extension to these approaches. The source code is available at https://anonymous.4open.science/r/ICCV-2023-Submission-PTUnifier/ and will be released in the final version of this paper.
****
A Large-scale Study of Spatiotemporal Representation Learning with a New Benchmark on Action Recognition
Andong Deng, Taojiannan Yang, Chen Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20519-20531
The goal of building a benchmark (suite of datasets) is to provide a unified protocol for fair evaluation and thus facilitate the evolution of a specific area. Nonetheless, we point out that existing protocols of action recognition could yield partial evaluations due to several limitations. To comprehensively probe the effectiveness of spatiotemporal representation learning, we introduce BEAR, a new BEnchmark on video Action Recognition. BEAR is a collection of 18 video datasets grouped into 5 categories (anomaly, gesture, daily, sports, and instructional), which covers a diverse set of real-world applications. With BEAR, we thoroughly evaluate 6 common spatiotemporal models pre-trained by both supervised and self-supervised learning. We also report transfer performance via standard finetuning, few-shot fine-tuning, and unsupervised domain adaptation. Our observation suggests that the current state-of-the-art cannot solidly guarantee high performance on datasets close to real-world applications, and we hope BEAR can serve as a fair and challenging evaluation benchmark to gain insights on building next-generation spatiotemporal learners. Our dataset, code, and models are released at: https://github.com/AndongDeng/BEAR
****
Video Background Music Generation: Dataset, Method and Evaluation
Le Zhuo, Zhaokai Wang, Baisen Wang, Yue Liao, Chenxi Bao, Stanley Peng, Songhao Han, Aixi Zhang, Fei Fang, Si Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15637-15647
Music is essential when editing videos, but selecting music manually is difficult and time-consuming. Thus, we seek to automatically generate background music tracks given video input. This is a challenging task since it requires music-video datasets, efficient architectures for video-to-music generation, and reasonable metrics, none of which currently exist. To close this gap, we introduce a complete recipe including dataset, benchmark model, and evaluation metric for video background music generation. We present SymMV, a video and symbolic music datase

t with various musical annotations. To the best of our knowledge, it is the first video-music dataset with rich musical annotations. We also propose a benchmark video background music generation framework named V-MusProd, which utilizes music priors of chords, melody, and accompaniment along with video-music relations of semantic, color, and motion features. To address the lack of objective metrics for video-music correspondence, we design a retrieval-based metric VMCP built upon a powerful video-music representation learning model. Experiments show that with our dataset, V-MusProd outperforms the state-of-the-art method in both music quality and correspondence with videos. We believe our dataset, benchmark model, and evaluation metric will boost the development of video background music generation. Our dataset and code are available at https://github.com/zhuole1025/SymMV.
****

HoloFusion: Towards Photo-realistic 3D Generative Modeling

Animesh Karnewar, Niloy J. Mitra, Andrea Vedaldi, David Novotny; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22976-22985

Diffusion-based image generators can now produce high-quality and diverse samples, but their success has yet to fully translate to 3D generation: existing diffusion methods can either generate low-resolution but 3D consistent outputs, or detailed 2D views of 3D objects with potential structural defects and lacking either view consistency or realism. We present HoloFusion, a method that combines the best of these approaches to produce high-fidelity, plausible, and diverse 3D samples while learning from a collection of multi-view 2D images only. The method first generates coarse 3D samples using a variant of the recently proposed HoloDiffusion generator. Then, it independently renders and upsamples a large number of views of the coarse 3D model, super-resolves them to add detail, and distills those into a single, high-fidelity implicit 3D representation, which also ensures view-consistency of the final renders. The super-resolution network is trained as an integral part of HoloFusion, and the final distillation uses a new sampling scheme to capture the space of super-resolved signals. We compare our method against existing baselines, including DreamFusion, Get3D, EG3D, and HoloDiffusion, and achieve, to the best of our knowledge, the most realistic results on the challenging CO3Dv2 dataset.
****

ProtoTransfer: Cross-Modal Prototype Transfer for Point Cloud Segmentation

Pin Tang, Hai-Ming Xu, Chao Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3337-3347

Knowledge transfer from multi-modal, i.e., LiDAR points and images, to a single LiDAR modal can take advantage of complimentary information from modal-fusion but keep a single modal inference speed, showing a promising direction for point cloud semantic segmentation in autonomous driving. Recent advances in point cloud segmentation distill knowledge from strictly aligned point-pixel fusion features while leaving a large number of unmatched image pixels unexplored and unmatched LiDAR points under-benefited. In this paper, we propose a novel approach, named ProtoTransfer, which not only fully exploits image representations but also transfers the learned multi-modal knowledge to all point cloud features. Specifically, based on the basic multi-modal learning framework, we build up a class-wise prototype bank from the strictly-aligned fusion features and encourage all the point cloud features to learn from the prototypes during model training. Moreover, to exploit the massive unmatched point and pixel features, we use a pseudo-labeling scheme and further accumulate these features into the class-wise prototype bank with a carefully designed fusion strategy. Without bells and whistles, our approach demonstrates superior performance over the published state-of-the-arts on two large-scale benchmarks, i.e., nuScenes and SemanticKITTI, and ranks 2nd on the competitive nuScenes Lidarseg challenge leaderboard.
****

Improving Continuous Sign Language Recognition with Cross-Lingual Signs

Fangyun Wei, Yutong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23612-23621

This work dedicates to continuous sign language recognition (CSLR), which is a w
eakly supervised task dealing with the recognition of continuous signs from vide
os, without any prior knowledge about the temporal boundaries between consecutiv
e signs. Data scarcity heavily impedes the progress of CSLR. Existing approaches
 typically train CSLR models on a monolingual corpus, which is orders of magnitu
de smaller than that of speech recognition. In this work, we explore the feasibi
lity of utilizing multilingual sign language corpora to facilitate monolingual C
SLR. Our work is built upon the observation of cross-lingual signs, which origin
ate from different sign languages but have similar visual signals (e.g., hand sh
ape and motion). The underlying idea of our approach is to identify the cross-li
ngual signs in one sign language and properly leverage them as auxiliary trainin
g data to improve the recognition capability of another. To achieve the goal, we
 first build two sign language dictionaries containing isolated signs that appea
r in two datasets. Then we identify the sign-to-sign mappings between two sign l
anguages via a well-optimized isolated sign language recognition model. At last,
 we train a CSLR model on the combination of the target data with original label
s and the auxiliary data with mapped labels. Experimentally, our approach achiev
es state-of-the-art performance on two widely-used CSLR datasets: Phoenix-2014 a
nd Phoenix-2014T.
****
Markov Game Video Augmentation for Action Segmentation
Nicolas Aziere, Sinisa Todorovic; Proceedings of the IEEE/CVF International Conf
erence on Computer Vision (ICCV), 2023, pp. 13505-13514
This paper addresses data augmentation for action segmentation. Our key novelty
is that we augment the original training videos in the deep feature space, not i
n the visual spatiotemporal domain as done by previous work. For augmentation, w
e modify original deep features of video frames such that the resulting embeddin
gs fall closer to the class decision boundaries. Also, we edit action sequences
of the original training videos (a.k.a. transcripts) by inserting, deleting, and
 replacing actions such that the resulting transcripts are close in edit distanc
e to the ground truth ones. For our data augmentation we resort to reinforcement
 learning, instead of more common supervised learning, since we do not have acce
ss to reliable oracles which would provide supervision about the optimal data mo
difications in the deep feature space. For modifying frame embeddings, we use a
meta-model formulated as a Markov Game with multiple self-interested agents. Als
o, new transcripts are generated using a fast, parameter-free Monte Carlo tree s
earch. Our experiments show that the proposed data augmentation of the Breakfast
, GTEA, and 50Salads datasets leads to significant performance gains of several
state of the art action segmenters.
****
Deep Image Harmonization with Globally Guided Feature Transformation and Relatio
n Distillation
Li Niu, Linfeng Tan, Xinhao Tao, Junyan Cao, Fengjun Guo, Teng Long, Liqing Zhan
g; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 7723-7732
Given a composite image, image harmonization aims to adjust the foreground illum
ination to be consistent with background. Previous methods have explored transfo
rming foreground features to achieve competitive performance. In this work, we s
how that using global information to guide foreground feature transformation cou
ld achieve significant improvement. Besides, we propose to transfer the foregrou
nd-background relation from real images to composite images, which can provide i
ntermediate supervision for the transformed encoder features. Additionally, cons
idering the drawbacks of existing harmonization datasets, we also contribute a c
cHarmony dataset which simulates the natural illumination variation. Extensive e
xperiments on iHarmony4 and our contributed dataset demonstrate the superiority
of our method. Our ccHarmony dataset is released at https://github.com/bcmi/Imag
e-Harmonization-Dataset-ccHarmony.
****
TransIFF: An Instance-Level Feature Fusion Framework for Vehicle-Infrastructure
Cooperative 3D Detection with Transformers

Ziming Chen, Yifeng Shi, Jinrang Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18205-18214
Cooperation between vehicles and infrastructure is vital to enhancing the safety of autonomous driving. Two significant and contradictory challenges now stand in the collaborative perception: fusion accuracy and communication bandwidth. Previous intermediate fusion methods that transmit features balance the accuracy and bandwidth compared with early fusion and late fusion, but usually have problems with feature alignment and domain gaps, and the bandwidth usage still falls short of the industrial application standard to our best knowledge. In this paper, we propose TransIFF, an instance-level feature fusion framework with transformers that can effectively reduce bandwidth usage. Furthermore, it can align the domain gaps between vehicle and infrastructure features, and improve the robustness of feature fusion, leading to a high cooperative perception accuracy. TransIFF is composed of three components: a vehicle-side network, an infrastructure-side network, and a vehicle-infrastructure fusion network. Initially, the vehicle-side and infrastructure-side networks independently generate instance-level features. Subsequently, the infrastructure-side instance-level features are transmitted to the vehicles, significantly reducing the communication bandwidth usage. Finally, in the vehicle-infrastructure fusion network, Cross-Domain Adaptation (CDA) module is designed to align the feature domains, followed by Feature Magnet (FM) module which can adaptively fuse the instance features and achieve a robust feature fusion. TransIFF yields state-of-the-art performance on the widely used real-world vehicle-infrastructure cooperative benchmark DAIR-V2X, achieving 59.62% AP with only 2^12 bytes bandwidth consumption.
****
RegFormer: An Efficient Projection-Aware Transformer Network for Large-Scale Point Cloud Registration
Jiuming Liu, Guangming Wang, Zhe Liu, Chaokang Jiang, Marc Pollefeys, Hesheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8451-8460
Although point cloud registration has achieved remarkable advances in object-level and indoor scenes, large-scale registration methods are rarely explored. Challenges mainly arise from the huge point number, complex distribution, and outliers of outdoor LiDAR scans. In addition, most existing registration works generally adopt a two-stage paradigm: They first find correspondences by extracting discriminative local features and then leverage estimators (eg. RANSAC) to filter outliers, which are highly dependent on well-designed descriptors and post-processing choices. To address these problems, we propose an end-to-end transformer network (RegFormer) for large-scale point cloud alignment without any further post-processing. Specifically, a projection-aware hierarchical transformer is proposed to capture long-range dependencies and filter outliers by extracting point features globally. Our transformer has linear complexity, which guarantees high efficiency even for large-scale scenes. Furthermore, to effectively reduce mismatches, a bijective association transformer is designed for regressing the initial transformation. Extensive experiments on KITTI and NuScenes datasets demonstrate that our RegFormer achieves competitive performance in terms of both accuracy and efficiency. Codes are available at https://github.com/IRMVLab/RegFormer.
****
Masked Retraining Teacher-Student Framework for Domain Adaptive Object Detection
Zijing Zhao, Sitong Wei, Qingchao Chen, Dehui Li, Yifan Yang, Yuxin Peng, Yang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19039-19049
Domain adaptive Object Detection (DAOD) leverages a labeled domain (source) to learn an object detector generalizing to a novel domain without annotation (target). Recent advances use a teacher-student framework, i.e., a student model is supervised by the pseudo labels from a teacher model. Though great success, they suffer from the limited number of pseudo boxes with incorrect predictions caused by the domain shift, misleading the student model to get sub-optimal results. To mitigate this problem, we propose Masked Retraining Teacher-student framework (MRT) which leverages masked autoencoder and selective retraining mechanism on de

tection transformer. Specifically, we present a customized design of masked auto
encoder branch, masking the multi-scale feature maps of target images and recons
tructing features by the encoder of the student model and an auxiliary decoder.
This helps the student model capture target domain characteristics and become a
more data-efficient learner to gain knowledge from the limited number of pseudo
boxes. Furthermore, we adopt selective retraining mechanism, periodically re-ini
tializing certain parts of the student parameters with masked autoencoder refine
d weights to allow the model to jump out of the local optimum biased to the inco
rrect pseudo labels. Experimental results on three DAOD benchmarks demonstrate t
he effectiveness of our method. Code can be found at https://github.com/JeremyZh
ao1998/MRT-release.
****
Prune Spatio-temporal Tokens by Semantic-aware Temporal Accumulation
Shuangrui Ding, Peisen Zhao, Xiaopeng Zhang, Rui Qian, Hongkai Xiong, Qi Tian; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 16945-16956
Transformers have become the primary backbone of the computer vision community d
ue to their impressive performance. However, the unfriendly computation cost imp
edes their potential in the video recognition domain. To optimize the speed-accu
racy trade-off, we propose Semantic-aware Temporal Accumulation score (STA) to p
rune spatio-temporal tokens integrally. STA score considers two critical factors
: temporal redundancy and semantic importance. The former depicts a specific reg
ion based on whether it is a new occurrence or a seen entity by aggregating toke
n-to-token similarity in consecutive frames while the latter evaluates each toke
n based on its contribution to the overall prediction. As a result, tokens with
higher scores of STA carry more temporal redundancy as well as lower semantics t
hus being pruned. Based on the STA score, we are able to progressively prune the
tokens without introducing any additional parameters or requiring further re-tr
aining. We directly apply the STA module to off-the-shelf ViT and VideoSwin back
bones, and the empirical results on Kinetics-400 and Something-Something V2 achi
eve over 30% computation reduction with a negligible 0.2% accuracy drop. The co
de is released at https://github.com/Mark12Ding/STA.
****
VQ3D: Learning a 3D-Aware Generative Model on ImageNet
Kyle Sargent, Jing Yu Koh, Han Zhang, Huiwen Chang, Charles Herrmann, Pratul Sri
nivasan, Jiajun Wu, Deqing Sun; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 4240-4250
Recent work has shown the possibility of training generative models of 3D conten
t from 2D image collections on small datasets corresponding to a single object c
lass, such as human faces, animal faces, or cars. However, these models struggle
on larger, more complex datasets. To model diverse and unconstrained image coll
ections such as ImageNet, we present VQ3D, which introduces a NeRF-based decoder
into a two-stage vector-quantized autoencoder. Our Stage 1 allows for the recon
struction of an input image and the ability to change the camera position around
the image, and our Stage 2 allows for the generation of new 3D scenes. VQ3D is
capable of generating and reconstructing 3D-aware images from the 1000-class Ima
geNet dataset of 1.2 million training images, and achieves a competitive ImageNe
t generation FID score of 16.8.
****
Growing a Brain with Sparsity-Inducing Generation for Continual Learning
Hyundong Jin, Gyeong-hyeon Kim, Chanho Ahn, Eunwoo Kim; Proceedings of the IEEE/
CVF International Conference on Computer Vision (ICCV), 2023, pp. 18961-18970
Deep neural networks suffer from catastrophic forgetting in continual learning,
where they tend to lose information about previously learned tasks when optimizi
ng a new incoming task. Recent strategies isolate the important parameters for p
revious tasks to retain old knowledge while learning the new task. However, usin
g the fixed old knowledge might act as an obstacle to capturing novel representa
tions. To overcome this limitation, we propose a framework that evolves the prev
iously allocated parameters by absorbing the knowledge of the new task. The appr
oach performs under two different networks. The base network learns knowledge of

sequential tasks, and the sparsity-inducing hypernetwork generates parameters for each time step for evolving old knowledge. The generated parameters transform old parameters of the base network to reflect the new knowledge. We design the hypernetwork to generate sparse parameters conditional to the task-specific information and the structural information of the base network. We evaluate the proposed approach on class-incremental and task-incremental learning scenarios for image classification and video action recognition tasks. Experimental results show that the proposed method consistently outperforms a large variety of continual learning approaches for those scenarios by evolving old knowledge.
****
Cross-Ray Neural Radiance Fields for Novel-View Synthesis from Unconstrained Image Collections
Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, Mingkui Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15901-15911
Neural Radiance Fields (NeRF) is a revolutionary approach for rendering scenes by sampling a single ray per pixel and it has demonstrated impressive capabilities in novel-view synthesis from static scene images. However, in practice, we usually need to recover NeRF from unconstrained image collections, which poses two challenges: 1) the images often have dynamic changes in appearance because of different capturing time and camera settings; 2) the images may contain transient objects such as humans and cars, leading to occlusion and ghosting artifacts. Conventional approaches seek to address these challenges by locally utilizing a single ray to synthesize a color of a pixel. In contrast, humans typically perceive appearance and objects by globally utilizing information across multiple pixels. To mimic the perception process of humans, in this paper, we propose Cross-Ray NeRF (CR-NeRF) that leverages interactive information across multiple rays to synthesize occlusion-free novel views with the same appearances as the images. Specifically, to model varying appearances, we first propose to represent multiple rays with a novel cross-ray feature and then recover the appearance by fusing global statistics, i.e., feature covariance of the rays and the image appearance. Moreover, to avoid occlusion introduced by transient objects, we propose a transient objects handler and introduce a grid sampling strategy for masking out the transient objects. We theoretically find that leveraging correlation across multiple rays promotes capturing more global information. Moreover, extensive experimental results on large real-world datasets verify the effectiveness of CR-NeRF.
****
Graphics2RAW: Mapping Computer Graphics Images to Sensor RAW Images
Donghwan Seo, Abhijith Punnappurath, Luxi Zhao, Abdelrahman Abdelhamed, Sai Kiran Tedla, Sanguk Park , Jihwan Choe, Michael S. Brown; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12622-12631
Computer graphics (CG) rendering platforms produce imagery with ever-increasing photo realism. The narrowing domain gap between real and synthetic imagery makes it possible to use CG images as training data for deep learning models targeting high-level computer vision tasks, such as autonomous driving and semantic segmentation. CG images, however, are currently not suitable for low-level vision tasks targeting RAW sensor images. This is because RAW images are encoded in sensor-specific color spaces and incur pre-white-balance color casts caused by the sensor's response to scene illumination. CG images are rendered directly to a device-independent perceptual color space without needing white balancing. As a result, it is necessary to apply a mapping procedure to close the domain gap between graphics and RAW images. To this end, we introduce a framework to process graphics images to mimic RAW sensor images accurately. Our approach allows a one-to-many mapping, where a single graphics image can be transformed to match multiple sensors and multiple scene illuminations. In addition, our approach requires only a handful of example RAW-DNG files from the target sensor as parameters for the mapping process. We compare our method to alternative strategies and show that our approach produces more realistic RAW images and provides better results on three low-level vision tasks: RAW denoising, illumination estimation, and neural

rendering for night photography. Finally, as part of this work, we provide a dataset of 292 realistic CG images for training low-light imaging models.
****
SPACE: Speech-driven Portrait Animation with Controllable Expression
Siddharth Gururani, Arun Mallya, Ting-Chun Wang, Rafael Valle, Ming-Yu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20914-20923
Animating portraits using speech has received growing attention in recent years, with various creative and practical use cases. An ideal generated video should have good lip sync with the audio, natural facial expressions and head motions, and high frame quality. In this work, we present SPACE, which uses speech and a single image to generate high-resolution, and expressive videos with realistic head pose, without requiring a driving video. It uses a multi-stage approach, combining the controllability of facial landmarks with the high-quality synthesis power of a pretrained face generator. SPACE also allows for the control of emotions and their intensities. Our method outperforms prior methods in objective metrics for image quality and facial motions and is strongly preferred by users in pair-wise comparisons. Please visit the project page to view the videos and to see more results: https://research.nvidia.com/labs/dir/space.
****
2D-3D Interlaced Transformer for Point Cloud Segmentation with Scene-Level Supervision
Cheng-Kun Yang, Min-Hung Chen, Yung-Yu Chuang, Yen-Yu Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 977-987
We present a Multimodal Interlaced Transformer (MIT) that jointly considers 2D and 3D data for weakly supervised point cloud segmentation. Research studies have shown that 2D and 3D features are complementary for point cloud segmentation. However, existing methods require extra 2D annotations to achieve 2D-3D information fusion. Considering the high annotation cost of point clouds, effective 2D and 3D feature fusion based on weakly supervised learning is in great demand. To this end, we propose a transformer model with two encoders and one decoder for weakly supervised point cloud segmentation using only scene-level class tags. Specifically, the two encoders compute the self-attended features for 3D point clouds and 2D multi-view images, respectively. The decoder implements interlaced 2D-3D cross-attention and carries out implicit 2D and 3D feature fusion. We alternately switch the roles of queries and key-value pairs in the decoder layers. It turns out that the 2D and 3D features are iteratively enriched by each other. Experiments show that it performs favorably against existing weakly supervised point cloud segmentation methods by a large margin on the S3DIS and ScanNet benchmarks.
****
Collecting The Puzzle Pieces: Disentangled Self-Driven Human Pose Transfer by Permuting Textures
Nannan Li, Kevin J Shih, Bryan A. Plummer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7126-7137
Human pose transfer synthesizes new view(s) of a person for a given pose. Recent work achieves this via self-reconstruction, which disentangles a person's pose and texture information by breaking down the person into several parts, then recombines them to reconstruct the person. However, this part-level disentanglement preserves some pose information that can create unwanted artifacts. In this paper, we propose Pose Transfer by Permuting Textures, a self-driven human pose transfer approach that disentangles pose from texture at the patch-level. Specifically, we remove pose from an input image by permuting image patches so only texture information remains. Then we reconstruct the input image by sampling from the permuted textures to achieve patch-level disentanglement. To reduce the noise and recover clothing shape information from the permuted patches, we employ encoders with multiple kernel sizes in a triple branch network. Extensive experiments on DeepFashion and Market-1501 show that our model improves the quality of generated images in terms of FID, LPIPS and SSIM over other self-driven methods, and even outperforming some fully-supervised methods. A user study also shows that

among self-driven approaches, images generated by our method are preferred in 68 % of cases over prior work. Code is available at https://github.com/NannanLi999/pt_square.
****
VAD: Vectorized Scene Representation for Efficient Autonomous Driving
Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, Xinggang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8340-8350
Autonomous driving requires a comprehensive understanding of the surrounding environment for reliable trajectory planning. Previous works rely on dense rasterized scene representation (e.g., agent occupancy and semantic map) to perform planning, which is computationally intensive and misses the instance-level structure information. In this paper, we propose VAD, an end-to-end vectorized paradigm for autonomous driving, which models the driving scene as a fully vectorized representation. The proposed vectorized paradigm has two significant advantages. On one hand, VAD exploits the vectorized agent motion and map elements as explicit instance-level planning constraints which effectively improves planning safety. On the other hand, VAD runs much faster than previous end-to-end planning methods by getting rid of computation-intensive rasterized representation and hand-designed post-processing steps. VAD achieves state-of-the-art end-to-end planning performance on the nuScenes dataset, outperforming the previous best method by a large margin. Our base model, VAD-Base, greatly reduces the average collision rate by 29.0% and runs 2.5x faster. Besides, a lightweight variant, VAD-Tiny, greatly improves the inference speed (up to 9.3x) while achieving comparable planning performance. We believe the excellent performance and the high efficiency of VAD are critical for the real-world deployment of an autonomous driving system. Code and models are available at https://github.com/hustvl/VAD for facilitating future research.
****
End-to-end 3D Tracking with Decoupled Queries
Yanwei Li, Zhiding Yu, Jonah Philion, Anima Anandkumar, Sanja Fidler, Jiaya Jia, Jose Alvarez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18302-18311
In this work, we present an end-to-end framework for camera-based 3D multi-object tracking, called DQTrack. To avoid heuristic design in detection-based trackers, recent query-based approaches deal with identity-agnostic detection and identity-aware tracking in a single embedding. However, it brings inferior performance because of the inherent representation conflict. To address this issue, we decouple the single embedding into separated queries, i.e., object query and track query. Unlike previous detection-based and query-based methods, the decoupled-query paradigm utilizes task-specific queries and still maintains the compact pipeline without complex post-processing. Moreover, the learnable association and temporal update are designed to provide differentiable trajectory association and frame-by-frame query update, respectively. The proposed DQTrack is demonstrated to achieve consistent gains in various benchmarks, outperforming all previous tracking-by-detection and learning-based methods on the nuScenes dataset.
****
Sound Localization from Motion: Jointly Learning Sound Direction and Camera Rotation
Ziyang Chen, Shengyi Qian, Andrew Owens; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7897-7908
The images and sounds that we perceive undergo subtle but geometrically consistent changes as we rotate our heads. In this paper, we use these cues to solve a problem we call Sound Localization from Motion (SLfM): jointly estimating camera rotation and localizing sound sources. We learn to solve these tasks solely through self-supervision. A visual model predicts camera rotation from a pair of images, while an audio model predicts the direction of sound sources from binaural sounds. We train these models to generate predictions that agree with one another. At test time, the models can be deployed independently. To obtain a feature representation that is well-suited to solving this challenging problem, we also p

ropose a method for learning an audio-visual representation through cross-view binauralization: estimating binaural sound from one view, given images and sound from another. Our model can successfully estimate accurate rotations on both real and synthetic scenes, and localize sound sources with accuracy competitive with state-of-the-art self-supervised approaches.
****
Batch-based Model Registration for Fast 3D Sherd Reconstruction
Jiepeng Wang, Congyi Zhang, Peng Wang, Xin Li, Peter J. Cobb, Christian Theobalt, Wenping Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14519-14529
3D reconstruction techniques have widely been used for digital documentation of archaeological fragments. However, efficient digital capture of fragments remains as a challenge. In this work, we aim to develop a portable, high-throughput, and accurate reconstruction system for efficient digitization of fragments excavated in archaeological sites. To realize high-throughput digitization of large numbers of objects, an effective strategy is to perform scanning and reconstruction in batches. However, effective batch-based scanning and reconstruction face two key challenges: 1) how to correlate partial scans of the same object from multiple batch scans, and 2) how to register and reconstruct complete models from partial scans that exhibit only small overlaps. To tackle these two challenges, we develop a new batch-based matching algorithm that pairs the front and back sides of the fragments, and a new Bilateral Boundary ICP algorithm that can register partial scans sharing very narrow overlapping regions. Extensive validation in labs and testing in excavation sites demonstrate that these designs enable efficient batch-based scanning for fragments. We show that such a batch-based scanning and reconstruction pipeline can have immediate applications on digitizing sherds in archaeological excavations.
****
HiFace: High-Fidelity 3D Face Reconstruction by Learning Static and Dynamic Details
Zenghao Chai, Tianke Zhang, Tianyu He, Xu Tan, Tadas Baltrusaitis, HsiangTao Wu, Runnan Li, Sheng Zhao, Chun Yuan, Jiang Bian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9087-9098
3D Morphable Models (3DMMs) demonstrate great potential for reconstructing faithful and animatable 3D facial surfaces from a single image. The facial surface is influenced by the coarse shape, as well as the static detail (e,g., person-specific appearance) and dynamic detail (e.g., expression-driven wrinkles). Previous work struggles to decouple the static and dynamic details through image-level supervision, leading to reconstructions that are not realistic. In this paper, we aim at high-fidelity 3D face reconstruction and propose HiFace to explicitly model the static and dynamic details. Specifically, the static detail is modeled as the linear combination of a displacement basis, while the dynamic detail is modeled as the linear interpolation of two displacement maps with polarized expressions. We exploit several loss functions to jointly learn the coarse shape and fine details with both synthetic and real-world datasets, which enable HiFace to reconstruct high-fidelity 3D shapes with animatable details. Extensive quantitative and qualitative experiments demonstrate that HiFace presents state-of-the-art reconstruction quality and faithfully recovers both the static and dynamic details.
****
Fast and Accurate Transferability Measurement by Evaluating Intra-class Feature Variance
Huiwen Xu, U Kang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11474-11482
Given a set of pre-trained models, how can we quickly and accurately find the most useful pre-trained model for a downstream task? Transferability measurement is to quantify how transferable is a pre-trained model learned on a source task to a target task. It is used for quickly ranking pre-trained models for a given task and thus becomes a crucial step for transfer learning. Existing methods measure transferability as the discrimination ability of a source model for a target

data before transfer learning, which cannot accurately estimate the fine-tuning performance. Some of them restrict the application of transferability measurement in selecting the best supervised pre-trained models that have classifiers. It is important to have a general method for measuring transferability that can be applied in a variety of situations, such as selecting the best self-supervised pre-trained models that do not have classifiers, and selecting the best transferring layer for a target task.
In this work, we propose TMI (TRANSFERABILITY MEASUREMENT WITH INTRA-CLASS FEATURE VARIANCE), a fast and accurate algorithm to measure transferability. We view transferability as the generalization of a pre-trained model on a target task by measuring intra-class feature variance. Intra-class variance evaluates the adaptability of the model to a new task, which measures how transferable the model is. Compared to previous studies that estimate how discriminative the models are, intra-class variance is more accurate than those as it does not require an optimal feature extractor and classifier. Extensive experiments on real-world datasets show that TMI outperforms competitors for selecting the top-5 best models, and exhibits consistently better correlation in 13 out of 17 cases.
****
Deformable Model-Driven Neural Rendering for High-Fidelity 3D Reconstruction of Human Heads Under Low-View Settings
Baixin Xu, Jiarui Zhang, Kwan-Yee Lin, Chen Qian, Ying He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17924-17934
Reconstructing 3D human heads in low-view settings presents technical challenges, mainly due to the pronounced risk of overfitting with limited views and high-frequency signals. To address this, we propose geometry decomposition and adopt a two-stage, coarse-to-fine training strategy, allowing for progressively capturing high-frequency geometric details. We represent 3D human heads using the zero level-set of a combined signed distance field, comprising a smooth template, a non-rigid deformation, and a high-frequency displacement field. The template captures features that are independent of both identity and expression and is co-trained with the deformation network across multiple individuals with sparse and randomly selected views. The displacement field, capturing individual-specific details, undergoes separate training for each person. Our network training does not require 3D supervision or object masks. Experimental results demonstrate the effectiveness and robustness of our geometry decomposition and two-stage training strategy. Our method outperforms existing neural rendering approaches in terms of reconstruction accuracy and novel view synthesis under low-view settings. Moreover, the pre-trained template serves a good initialization for our model when encountering unseen individuals.
****
Algebraically Rigorous Quaternion Framework for the Neural Network Pose Estimation Problem
Chen Lin, Andrew J. Hanson, Sonya M. Hanson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14097-14106
The 3D pose estimation problem -- aligning pairs of noisy 3D point clouds -- is a problem with a wide variety of real-world applications. Here we focus on the use of quaternion-based neural network approaches to this problem and apparent anomalies that have arisen in previous efforts to resolve them. In addressing these anomalies, we draw heavily from the extensive literature on closed-form methods to solve this problem. We suggest that the major concerns that have been put forward could be resolved using a simple multi-valued training target derived from rigorous theoretical properties of the rotation-to-quaternion map of Bar-Itzhack. This multi-valued training target is then demonstrated to have good performance for both simulated and ModelNet targets. We provide a comprehensive theoretical context, using the quaternion adjugate, to confirm and establish the necessity of replacing single-valued quaternion functions by quaternions treated in the extended domain of multiple-charted manifolds.
****
Prompt Tuning Inversion for Text-driven Image Editing Using Diffusion Models
Wenkai Dong, Song Xue, Xiaoyue Duan, Shumin Han; Proceedings of the IEEE/CVF Int

ernational Conference on Computer Vision (ICCV), 2023, pp. 7430-7440
Recently large-scale language-image models (e.g., text-guided diffusion models) have considerably improved the image generation capabilities to generate photorealistic images in various domains. Based on this success, current image editing methods use texts to achieve intuitive and versatile modification of images. To edit a real image using diffusion models, one must first invert the image to a noisy latent from which an edited image is sampled with a target text prompt. However, most methods lack one of the following: user-friendliness (e.g., additional masks or precise descriptions of the input image are required), generalization to larger domains, or high fidelity to the input image. In this paper, we design an accurate and quick inversion technique, Prompt Tuning Inversion, for text-driven image editing. Specifically, our proposed editing method consists of a reconstruction stage and an editing stage. In the first stage, we encode the information of the input image into a learnable conditional embedding via Prompt Tuning Inversion. In the second stage, we apply classifier-free guidance to sample the edited image, where the conditional embedding is calculated by linearly interpolating between the target embedding and the optimized one obtained in the first stage. This technique ensures a superior trade-off between editability and high fidelity to the input image of our method. For example, we can change the color of a specific object while preserving its original shape and background under the guidance of only a target text prompt. Extensive experiments on ImageNet demonstrate the superior editing performance of our method compared to the state-of-the-art baselines.
****
CVSformer: Cross-View Synthesis Transformer for Semantic Scene Completion
Haotian Dong, Enhui Ma, Lubo Wang, Miaohui Wang, Wuyuan Xie, Qing Guo, Ping Li, Lingyu Liang, Kairui Yang, Di Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8874-8883
Semantic scene completion (SSC) requires an accurate understanding of the geometric and semantic relationships between the objects in the 3D scene for reasoning the occluded objects. The popular SSC methods voxelize the 3D objects, allowing the deep 3D convolutional network (3D CNN) to learn the object relationships from the complex scenes. However, the current networks lack the controllable kernels to model the object relationship across multiple views, where appropriate views provide the relevant information for suggesting the existence of the occluded objects. In this paper, we propose Cross-View Synthesis Transformer (CVSformer), which consists of Multi-View Feature Synthesis and Cross-View Transformer for learning cross-view object relationships. In the multi-view feature synthesis, we use a set of 3D convolutional kernels rotated differently to compute the multi-view features for each voxel. In the cross-view transformer, we employ the cross-view fusion to comprehensively learn the cross-view relationships, which form useful information for enhancing the features of individual views. We use the enhanced features to predict the geometric occupancies and semantic labels of all voxels. We evaluate CVSformer on public datasets, where CVSformer yields state-of-the-art results. Our code is available at https://github.com/donghaotian123/CVSformer.
****
UrbanGIRAFFE: Representing Urban Scenes as Compositional Generative Neural Feature Fields
Yuanbo Yang, Yifei Yang, Hanlei Guo, Rong Xiong, Yue Wang, Yiyi Liao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9199-9210
Generating photorealistic images with controllable camera pose and scene contents is essential for many applications including AR/VR and simulation. Despite the fact that rapid progress has been made in 3D-aware generative models, most existing methods focus on object-centric images and are not applicable to generating urban scenes for free camera viewpoint control and scene editing. To address this challenging task, we propose UrbanGIRAFFE, which uses a coarse 3D panoptic prior, including the layout distribution of uncountable stuff and countable objects, to provide semantic and geometric prior. Our model is compositional and contr

ollable as it breaks down the scene into stuff, objects, and sky. Using stuff pr ior in the form of semantic voxel grids, we build a conditioned stuff generator that effectively incorporates the coarse semantic and geometry information. The object layout prior further allows us to learn an object generator from cluttere d scenes. With proper loss functions, our approach facilitates photorealistic 3D -aware image synthesis with diverse controllability, including large camera move ment, stuff editing, and object manipulation. We validate the effectiveness of o ur model on both synthetic and real-world datasets, including the challenging KI TTI-360 dataset.
****
UnitedHuman: Harnessing Multi-Source Data for High-Resolution Human Generation
Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Wayne Wu, Ziwei Liu; Proceed ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, p p. 7301-7311
Human generation has achieved significant progress. Nonetheless, existing method s still struggle to synthesize specific regions such as faces and hands. We argu e that the main reason is rooted in the training data. A holistic human dataset inevitably has insufficient and low-resolution information on local parts. There fore, we propose to use multi-source datasets with various resolution images to jointly learn a high-resolution human generative model. However, multi-source da ta inherently a) contains different parts that do not spatially align into a coh erent human, and b) comes with different scales. To tackle these challenges, we propose an end-to-end framework, UnitedHuman, that empowers continuous GAN with the ability to effectively utilize multi-source data for high-resolution human g eneration. Specifically, 1) we design a Multi-Source Spatial Transformer that sp atially aligns multi-source images to full-body space with a human parametric mo del. 2) Next, a continuous GAN is proposed with global-structural guidance and C utMix consistency. Patches from different datasets are then sampled and transfor med to supervise the training of this scale-invariant generative model. Extensiv e experiments demonstrate that our model jointly learned from multi-source data achieves superior quality than those learned from a holistic dataset.
****
Active Neural Mapping
Zike Yan, Haoxiang Yang, Hongbin Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10981-10992
We address the problem of active mapping with a continually-learned neural scene representation, namely Active Neural Mapping. The key lies in actively finding the target space to be explored with efficient agent movement, thus minimizing t he map uncertainty on-the-fly within a previously unseen environment. In this pa per, we examine the weight space of the continually-learned neural field, and sh ow empirically that the neural variability, the prediction robustness against ra ndom weight perturbation, can be directly utilized to measure the instant uncert ainty of the neural map. Together with the continuous geometric information inhe rited in the neural map, the agent can be guided to find a traversable path to g radually gain knowledge of the environment. We present for the first time an onl ine active mapping system with a coordinate-based implicit neural representation . Experiments in the visually-realistic Gibson and Matterport3D environment demo nstrate the efficacy of the proposed method.
****
Density-invariant Features for Distant Point Cloud Registration
Quan Liu, Hongzi Zhu, Yunsong Zhou, Hongyang Li, Shan Chang, Minyi Guo; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp . 18215-18225
Registration of distant outdoor LiDAR point clouds is crucial to extending the 3 D vision of collaborative autonomous vehicles, and yet is challenging due to sma ll overlapping area and a huge disparity between observed point densities. In th is paper, we propose Group-wise Contrastive Learning (GCL) scheme to extract den sity-invariant geometric features to register distant outdoor LiDAR point clouds . We mark through theoretical analysis and experiments that, contrastive positiv es should be independent and identically distributed (i.i.d.), in order to train

density-invariant feature extractors. We propose upon the conclusion a simple y
et effective training scheme to force the feature of multiple point clouds in th
e same spatial location (referred to as positive groups) to be similar, which na
turally avoids the sampling bias introduced by a pair of point clouds to conform
 with the i.i.d. principle. The resulting fully-convolutional feature extractor
is more powerful and density-invariant than state-of-the-art methods, improving
the registration recall of distant scenarios on KITTI and nuScenes benchmarks by
 40.9% and 26.9%, respectively. Code is available at https://github.com/liuQuan9
8/GCL.
****
UniverSeg: Universal Medical Image Segmentation
Victor Ion Butoi, Jose Javier Gonzalez Ortiz, Tianyu Ma, Mert R. Sabuncu, John G
uttag, Adrian V. Dalca; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2023, pp. 21438-21451
While deep learning models have become the predominant method for medical image
segmentation, they are typically not capable of generalizing to unseen segmentat
ion tasks involving new anatomies, image modalities, or labels. Given a new segm
entation task, researchers generally have to train or fine-tune models. This is
time-consuming and poses a substantial barrier for clinical researchers, who oft
en lack the resources and expertise to train neural networks. We present UniverS
eg, a method for solving unseen medical segmentation tasks without additional tr
aining. Given a query image and an example set of image-label pairs that define
a new segmentation task, UniverSeg employs a new CrossBlock mechanism to produce
 accurate segmentation maps without additional training. To achieve generalizati
on to new tasks, we have gathered and standardized a collection of 53 open-acces
s medical segmentation datasets with over 22,000 scans, which we refer to as Meg
aMedical. We used this collection to train UniverSeg on a diverse set of anatomi
es and imaging modalities. We demonstrate that UniverSeg substantially outperfor
ms several related methods on unseen tasks, and thoroughly analyze and draw insi
ghts about important aspects of the proposed system. The UniverSeg source code a
nd model weights are freely available at https://universeg.csail.mit.edu.
****
RecRecNet: Rectangling Rectified Wide-Angle Images by Thin-Plate Spline Model an
d DoF-based Curriculum Learning
Kang Liao, Lang Nie, Chunyu Lin, Zishuo Zheng, Yao Zhao; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10800-10809
The wide-angle lens shows appealing applications in VR technologies, but it intr
oduces severe radial distortion into its captured image. To recover the realisti
c scene, previous works devote to rectifying the content of the wide-angle image
. However, such a rectification solution inevitably distorts the image boundary,
 which changes related geometric distributions and misleads the current vision p
erception models. In this work, we explore constructing a win-win representation
 on both content and boundary by contributing a new learning model, i.e., Rectan
gling Rectification Network (RecRecNet). In particular, we propose a thin-plate
spline (TPS) module to formulate the non-linear and non-rigid transformation for
 rectangling images. By learning the control points on the rectified image, our
model can flexibly warp the source structure to the target domain and achieves a
n end-to-end unsupervised deformation. To relieve the complexity of structure ap
proximation, we then inspire our RecRecNet to learn the gradual deformation rule
s with a DoF (Degree of Freedom)-based curriculum learning. By increasing the Do
F in each curriculum stage, namely, from similarity transformation (4-DoF) to ho
mography transformation (8-DoF), the network is capable of investigating more de
tailed deformations, offering fast convergence on the final rectangling task. Ex
periments show the superiority of our solution over the compared methods on both
 quantitative and qualitative evaluations. The code and dataset will be made ava
ilable.
****
Neural Microfacet Fields for Inverse Rendering
Alexander Mai, Dor Verbin, Falko Kuester, Sara Fridovich-Keil; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 408-418

We present Neural Microfacet Fields, a method for recovering materials, geometry (volumetric density), and environmental illumination from a collection of images of a scene. Our method applies a microfacet reflectance model within a volumetric setting by treating each sample along the ray as a surface, rather than an emitter. Using surface-based Monte Carlo rendering in a volumetric setting enables our method to perform inverse rendering efficiently and enjoy recent advances in volume rendering. Our approach obtains similar performance as state-of-the-art methods for novel view synthesis and outperforms prior work in inverse rendering, capturing high fidelity geometry and high frequency illumination details.
****
Understanding Self-attention Mechanism via Dynamical System Perspective
Zhongzhan Huang, Mingfu Liang, Jinghui Qin, Shanshan Zhong, Liang Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1412-1422
The self-attention mechanism (SAM) is widely used in various fields of artificial intelligence and has successfully boosted the performance of different models. However, current explanations of this mechanism are mainly based on intuitions and experiences, while there still lacks direct modeling for how the SAM helps performance. To mitigate this issue, in this paper, based on the dynamical system perspective of the residual neural network, we first show that the intrinsic stiffness phenomenon (SP) in the high-precision solution of ordinary differential equations (ODEs) also widely exists in high-performance neural networks (NN). Thus the ability of NN to measure SP at the feature level is necessary to obtain high performance and is an important factor in the difficulty of training NN. Similar to the adaptive step-size method which is effective in solving stiff ODEs, we show that the SAM is also a stiffness-aware step size adaptor that can enhance the model's representational ability to measure intrinsic SP by refining the estimation of stiffness information and generating adaptive attention values, which provides a new understanding about why and how the SAM can benefit the model performance. This novel perspective can also explain the lottery ticket hypothesis in SAM, design new quantitative metrics of representational ability, and inspire a new theoretic-inspired approach, StepNet. Extensive experiments on several popular benchmarks demonstrate that StepNet can extract fine-grained stiffness information and measure SP accurately, leading to significant improvements in various visual tasks.
****
Learning Versatile 3D Shape Generation with Improved Auto-regressive Models
Simian Luo, Xuelin Qian, Yanwei Fu, Yinda Zhang, Ying Tai, Zhenyu Zhang, Chengjie Wang, Xiangyang Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14139-14149
Auto-Regressive (AR) models have achieved impressive results in 2D image generation by modeling joint distributions in the grid space. While this approach has been extended to the 3D domain for powerful shape generation, it still has two limitations: expensive computations on volumetric grids and ambiguous auto-regressive order along grid dimensions. To overcome these limitations, we propose the Improved Auto-regressive Model (ImAM) for 3D shape generation, which applies discrete representation learning based on a latent vector instead of volumetric grids. Our approach not only reduces computational costs but also preserves essential geometric details by learning the joint distribution in a more tractable order. Moreover, thanks to the simplicity of our model architecture, we can naturally extend it from unconditional to conditional generation by concatenating various conditioning inputs, such as point clouds, categories, images, and texts. Extensive experiments demonstrate that ImAM can synthesize diverse and faithful shapes of multiple categories, achieving state-of-the-art performance.
****
DETA: Denoised Task Adaptation for Few-Shot Learning
Ji Zhang, Lianli Gao, Xu Luo, Hengtao Shen, Jingkuan Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11541-11551
Test-time task adaptation in few-shot learning aims to adapt a pre-trained task-agnostic model for capturing task-specific knowledge of the test task, rely only

on few-labeled support samples. Previous approaches generally focus on developing advanced algorithms to achieve the goal, while neglecting the inherent problems of the given support samples. In fact, with only a handful of samples available, the adverse effect of either the image noise (a.k.a. X-noise) or the label noise (a.k.a. Y-noise) from support samples can be severely amplified. To address this challenge, in this work we propose DEnoised Task Adaptation (DETA), a first, unified image- and label-denoising framework orthogonal to existing task adaptation approaches. Without extra supervision, DETA filters out task-irrelevant, noisy representations by taking advantage of both global visual information and local region details of support samples. On the challenging Meta-Dataset, DETA consistently improves the performance of a broad spectrum of baseline methods applied on various pre-trained models. Notably, by tackling the overlooked image noise in Meta-Dataset, DETA establishes new state-of-the-art results. Code is released at https://github.com/JimZAI/DETA.
****
DDG-Net: Discriminability-Driven Graph Network for Weakly-supervised Temporal Action Localization
Xiaojun Tang, Junsong Fan, Chuanchen Luo, Zhaoxiang Zhang, Man Zhang, Zongyuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6622-6632
Weakly-supervised temporal action localization (WTAL) is a practical yet challenging task. Due to large-scale datasets, most existing methods use a network pretrained in other datasets to extract features, which are not suitable enough for WTAL. To address this problem, researchers design several modules for feature enhancement, which improve the performance of the localization module, especially modeling the temporal relationship between snippets. However, all of them omit that ambiguous snippets deliver contradictory information, which would reduce the discriminability of linked snippets. Considering this phenomenon, we propose Discriminability-Driven Graph Network (DDG-Net), which explicitly models ambiguous snippets and discriminative snippets with well-designed connections, preventing the transmission of ambiguous information and enhancing the discriminability of snippet-level representations. Additionally, we propose feature consistency loss to prevent the assimilation of features and drive the graph convolution network to generate more discriminative representations. Extensive experiments on THUMOS14 and ActivityNet1.2 benchmarks demonstrate the effectiveness of DDG-Net, establishing new state-of-the-art results on both datasets. Source code is available at https://github.com/XiaojunTang22/ICCV2023-DDGNet.
****
Diffusion Models as Masked Autoencoders
Chen Wei, Karttikeya Mangalam, Po-Yao Huang, Yanghao Li, Haoqi Fan, Hu Xu, Huiyu Wang, Cihang Xie, Alan Yuille, Christoph Feichtenhofer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16284-16294
There has been a longstanding belief that generation can facilitate a true understanding of visual data. In line with this, we revisit generatively pre-training visual representations in light of recent interest in denoising diffusion models. While directly pre-training with diffusion models does not produce strong representations, we condition diffusion models on masked input and formulate diffusion models as masked autoencoders (DiffMAE). Our approach is capable of (i) serving as a strong initialization for downstream recognition tasks, (ii) conducting high-quality image inpainting, and (iii) being effortlessly extended to video where it produces state-of-the-art classification accuracy. We further perform a comprehensive study on the pros and cons of design choices and build connections between diffusion models and masked autoencoders.
****
Robust Frame-to-Frame Camera Rotation Estimation in Crowded Scenes
Fabien Delattre, David Dirnfeld, Phat Nguyen, Stephen K Scarano, Michael J Jones, Pedro Miraldo, Erik Learned-Miller; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9752-9762
We present an approach to estimating camera rotation in crowded, real-world scenes from handheld monocular video. While camera rotation estimation is a well-stu

died problem, no previous methods exhibit both high accuracy and acceptable speed in this setting. Because the setting is not addressed well by other datasets, we provide a new dataset and benchmark, with high-accuracy, rigorously verified ground truth, on 17 video sequences. Methods developed for wide baseline stereo (e.g., 5-point methods) perform poorly on monocular video. On the other hand, methods used in autonomous driving (e.g., SLAM) leverage specific sensor setups, specific motion models, or local optimization strategies (lagging batch processing) and do not generalize well to handheld video. Finally, for dynamic scenes, commonly used robustification techniques like RANSAC require large numbers of iterations, and become prohibitively slow. We introduce a novel generalization of the Hough transform on SO(3) to efficiently and robustly find the camera rotation most compatible with optical flow. Among comparably fast methods, ours reduces error by almost 50% over the next best, and is more accurate than any method, irrespective of speed. This represents a strong new performance point for crowded scenes, an important setting for computer vision. The code and the dataset are available at https://fabiendelattre.com/robust-rotation-estimation.
****

Bayesian Prompt Learning for Image-Language Model Generalization
Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor G. Turrisi da Costa, Cees G.M. Snoek, Georgios Tzimiropoulos, Brais Martinez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15237-15246

Foundational image-language models have generated considerable interest due to their efficient adaptation to downstream tasks by prompt learning. Prompt learning treats part of the language model input as trainable while freezing the rest, and optimizes an Empirical Risk Minimization objective. However, Empirical Risk Minimization is known to suffer from distributional shifts which hurt generalizability to prompts unseen during training. By leveraging the regularization ability of Bayesian methods, we frame prompt learning from the Bayesian perspective and formulate it as a variational inference problem. Our approach regularizes the prompt space, reduces overfitting to the seen prompts and improves the prompt generalization on unseen prompts. Our framework is implemented by modeling the input prompt space in a probabilistic manner, as an a priori distribution which makes our proposal compatible with prompt learning approaches that are unconditional or conditional on the image. We demonstrate empirically on 15 benchmarks that Bayesian prompt learning provides an appropriate coverage of the prompt space, prevents learning spurious features, and exploits transferable invariant features. This results in better generalization of unseen prompts, even across different datasets and domains. Code available at: https://github.com/saic-fi/Bayesian Prompt-Learning
****

One-Shot Recognition of Any Material Anywhere Using Contrastive Learning with Physics-Based Rendering
Manuel S. Drehwald, Sagi Eppel, Jolina Li, Han Hao, Alan Aspuru-Guzik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23524-23533

Visual recognition of materials and their states is essential for understanding the world, from determining whether food is cooked, metal is rusted, or a chemical reaction has occurred. However, current image recognition methods are limited to specific classes and properties and can't handle the vast number of material states in the world.

To address this, we present MatSim: the first dataset and benchmark for computer vision-based recognition of similarities and transitions between materials and textures, focusing on identifying any material under any conditions using one or a few examples. The dataset contains synthetic and natural images. Synthetic images were rendered using giant collections of textures, objects, and environments generated by computer graphics artists. We use mixtures and gradual transitions between materials to allow the system to learn cases with smooth transitions between states (like gradually cooked food). We also render images with materials inside transparent containers to support beverage and chemistry lab use cases.

We use this dataset to train a Siamese net that identifies the same material in different objects, mixtures, and environments. The descriptor generated by this net can be used to identify the states of materials and their subclasses using a single image.
We also present the first few-shot material recognition benchmark with natural images from a wide range of fields, including the state of foods and beverages, types of grounds, and many other use cases. We show that a net trained on the MatSim synthetic dataset outperforms state-of-the-art models like Clip on the benchmark and also achieves good results on other unsupervised material classification tasks. Dataset, generation code and trained models have been made available at: https://github.com/ZuseZ4/MatSim-Dataset-Generator-Scripts-And-Neural-net
****
DiLiGenT-Pi: Photometric Stereo for Planar Surfaces with Rich Details - Benchmark Dataset and Beyond
Feishi Wang, Jieji Ren, Heng Guo, Mingjun Ren, Boxin Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9477-9487
Photometric stereo aims to recover detailed surface shapes from images captured under varying illuminations. However, existing real-world datasets primarily focus on evaluating photometric stereo for general non-Lambertian reflectances and feature bulgy shapes that have a certain height. As shape detail recovery is the key strength of photometric stereo over other 3D reconstruction techniques, and the near-planar surfaces widely exist in cultural relics and manufacturing workpieces, we present a new real-world dataset DiLiGenT-Pi containing 30 near-planar scenes with rich surface details. This dataset enables us to evaluate recent photometric stereo methods specifically for their ability to estimate shape details under diverse materials and to identify open problems such as near-planar surface normal estimation from uncalibrated photometric stereo and surface detail recovery for translucent materials. To inspire future research, this dataset will open soruced at https://photometricstereo.github.io/diligentpi.html.
****
Rethinking Data Distillation: Do Not Overlook Calibration
Dongyao Zhu, Bowen Lei, Jie Zhang, Yanbo Fang, Yiqun Xie, Ruqi Zhang, Dongkuan Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4935-4945
Neural networks trained on distilled data often produce over-confident output and require correction by calibration methods. Existing calibration methods such as temperature scaling and mixup work well for networks trained on original large-scale data. However, we find that these methods fail to calibrate networks trained on data distilled from large source datasets. In this paper, we show that distilled data lead to networks that are not calibratable due to (i) a more concentrated distribution of the maximum logits and (ii) the loss of information that is semantically meaningful but unrelated to classification tasks. To address this problem, we propose Masked Temperature Scaling (MTS) and Masked Distillation Training (MDT) which mitigate the limitations of distilled data and achieve better calibration results while maintaining the efficiency of dataset distillation.
****
Accurate and Fast Compressed Video Captioning
Yaojie Shen, Xin Gu, Kai Xu, Heng Fan, Longyin Wen, Libo Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15558-15567
Existing video captioning approaches typically require to first sample video frames from a decoded video and then conduct a subsequent process (e.g., feature extraction and/or captioning model learning). In this pipeline, manual frame sampling may ignore key information in videos and thus degrade performance. Additionally, redundant information in the sampled frames may result in low efficiency in the inference of video captioning. Addressing this, we study video captioning from a different perspective in compressed domain, which brings multi-fold advantages over the existing pipeline: 1) Compared to raw images from the decoded video, the compressed video, consisting of I-frames, motion vectors and residuals, i

s highly distinguishable, which allows us to leverage the entire video for learning without manual sampling through a specialized model design; 2) The captioning model is more efficient in inference as smaller and less redundant information is processed. We propose a simple yet effective end-to-end transformer in the compressed domain for video captioning that enables learning from the compressed video for captioning. We show that even with a simple design, our method can achieve state-of-the-art performance on different benchmarks while running almost 2x faster than existing approaches. Code is available at https://github.com/acherstyx/CoCap.
****
Building Vision Transformers with Hierarchy Aware Feature Aggregation
Yongjie Chen, Hongmin Liu, Haoran Yin, Bin Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5908-5918
Thanks to the excellent global modeling capability of attention mechanisms, the Vision Transformer has achieved better results than ConvNet in many computer tasks. However, in generating hierarchical feature maps, the Transformer still adopts the ConvNet feature aggregation scheme. This leads to the problem that the semantic information of the grid area of image becomes confused after feature aggregation, making it difficult for attention to accurately model global relationships. To address this, we propose the Hierarchy Aware Feature Aggregation framework (HAFA). HAFA enhances the extraction of local features adaptively in shallow layers where semantic information is weak, while is able to aggregate patches with similar semantics in deep layers. The clear semantic information of the aggregated patches, enables the attention mechanism to more accurately model global information at the semantic level. Extensive experiments show that after using the HAFA framework, significant improvements have been achieved relative to the baseline models in image classification, object detection, and semantic segmentation tasks.
****
Visible-Infrared Person Re-Identification via Semantic Alignment and Affinity Inference
Xingye Fang, Yang Yang, Ying Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11270-11279
Visible-infrared person re-identification (VI-ReID) focuses on matching the pedestrian images of the same identity captured by different modality cameras. The part-based methods achieve great success by extracting fine-grained features from feature maps. But most existing part-based methods employ horizontal division to obtain part features suffering from misalignment caused by irregular pedestrian movements. Moreover, most current methods use Euclidean or cosine distance of the output features to measure the similarity without considering the pedestrian relationships. Misaligned part features and naive inference methods both limit the performance of existing works. We propose a Semantic Alignment and Affinity Inference framework (SAAI), which aims to align latent semantic part features with the learnable prototypes and improve inference with affinity information. Specifically, we first propose semantic-aligned feature learning that employs the similarity between pixel-wise features and learnable prototypes to aggregate the latent semantic part features. Then, we devise an affinity inference module to optimize the inference with pedestrian relationships. Comprehensive experimental results conducted on the SYSU-MM01 and RegDB datasets demonstrate the favorable performance of our SAAI framework. Our code will be released at https://github.com/xiaoye-hhh/SAAI.
****
SAL-ViT: Towards Latency Efficient Private Inference on ViT using Selective Attention Search with a Learnable Softmax Approximation
Yuke Zhang, Dake Chen, Souvik Kundu, Chenghao Li, Peter A. Beerel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5116-5125
Recently, private inference (PI) has addressed the rising concern over data and model privacy in machine learning inference as a service. However, existing PI frameworks suffer from high computational and communication overheads due to the

expensive multi-party computation (MPC) protocols, particularly for large models
 such as vision transformers (ViT). The majority of this overhead is due to the
encrypted softmax operation in each self-attention layer. In this work, we prese
nt SAL-ViT with two novel techniques to boost PI efficiency on ViTs. Our first t
echnique is a learnable PI-efficient approximation to softmax, namely, learnable
 2Quad (L2Q), that introduces learnable scaling and shifting parameters to the p
rior 2Quad softmax approximation, enabling improvement in accuracy. Then, given
our observation that external attention (EA) presents lower PI latency than wide
ly-adopted self-attention (SA) at the cost of accuracy, we present a selective a
ttention search (SAS) method to integrate the strength of EA and SA. Specificall
y, for a given lightweight EA ViT, we leverage a constrained optimization proced
ure to selectively search and replace EA modules with SA alternatives to maximiz
e the accuracy. Our extensive experiments show that our SAL-ViT can averagely ac
hieve 1.28x, 1.28x, 1.14x lower PI latency with 1.79%, 1.41%, and 2.08% higher a
ccuracy compared to the existing alternatives, on CIFAR-10, CIFAR-100, and Tiny-
ImageNet, respectively.
****
TIJO: Trigger Inversion with Joint Optimization for Defending Multimodal Backdoo
red Models
Indranil Sur, Karan Sikka, Matthew Walmer, Kaushik Koneripalli, Anirban Roy, Xia
o Lin, Ajay Divakaran, Susmit Jha; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2023, pp. 165-175
We present a Multimodal Backdoor defense technique TIJO (Trigger Inversion using
 Joint Optimization). Recently Walmer et al. demonstrated successful backdoor at
tacks on multimodal models for the Visual Question Answering task. Their dual-ke
y backdoor trigger is split across two modalities (image and text), such that th
e backdoor is activated if and only if the trigger is present in both modalities
. We propose TIJO that defends against dual-key attacks through a joint optimiza
tion that reverse-engineers the trigger in both the image and text modalities. T
his joint optimization is challenging in multimodal models due to the disconnect
ed nature of the visual pipeline which consists of an offline feature extractor,
 whose output is then fused with the text using a fusion module. The key insight
 enabling the joint optimization in TIJO is that the trigger inversion needs to
be carried out in the object detection box feature space as opposed to the pixel
 space. We demonstrate the effectiveness of our method on the TrojVQA benchmark,
 where TIJO improves upon the state-of-the-art unimodal methods from an AUC of 0
.6 to 0.92 on multimodal dual-key backdoors. Furthermore, our method also improv
es upon the unimodal baselines on unimodal backdoors. We also present detailed a
blation studies as well as qualitative results to provide insights into our algo
rithm such as the critical importance of overlaying the inverted feature trigger
s on all visual features during trigger inversion.
****
DG3D: Generating High Quality 3D Textured Shapes by Learning to Discriminate Mul
ti-Modal Diffusion-Renderings
Qi Zuo, Yafei Song, Jianfang Li, Lin Liu, Liefeng Bo; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2023, pp. 14575-14584
Many virtual reality applications require massive 3D content, which impels the n
eed for low-cost and efficient modeling tools in terms of quality and quantity.
In this paper, we present a Diffusion-augmented Generative model to generate hig
h-fidelity 3D textured meshes that can be directly used in modern graphics engin
es. Challenges in directly generating textured mesh arise from the instability a
nd texture incompleteness of a hybrid framework which contains conversion betwee
n 2D features and 3D space. To alleviate these difficulties, DG3D incorporates a
 diffusion-based augmentation module into the min-max game between the 3D tetrah
edral mesh generator and 2D renderings discriminators, which stabilizes network
optimization and prevents mode collapse in vanilla GANs. We also suggest using m
ulti-modal renderings in discrimination to further increase the aesthetics and c
ompleteness of generated textures. Extensive experiments on the public benchmark
 and real scans show that our proposed DG3D outperforms existing state-of-the-ar
t methods by a large margin, i.e., 5%   40% in FID-3D score and 5%   10% in geom

etry-related metrics. Code is available at
https://github.com/seakforzq/DG3D.
****
Improving Adversarial Robustness of Masked Autoencoders via Test-time Frequency-domain Prompting
Qidong Huang, Xiaoyi Dong, Dongdong Chen, Yinpeng Chen, Lu Yuan, Gang Hua, Weiming Zhang, Nenghai Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1600-1610
In this paper, we investigate the adversarial robustness of vision transformers that are equipped with BERT pretraining (e.g., BEiT, MAE). A surprising observation is that MAE has significantly worse adversarial robustness than other BERT pretraining methods. This observation drives us to rethink the basic differences between these BERT pretraining methods and how these differences affect the robustness against adversarial perturbations. Our empirical analysis reveals that the adversarial robustness of BERT pretraining is highly related to the reconstruction target, i.e., predicting the raw pixels of masked image patches will degrade more adversarial robustness of the model than predicting the semantic context, since it guides the model to concentrate more on medium-/high-frequency components of images. Based on our analysis, we provide a simple yet effective way to boost the adversarial robustness of MAE. The basic idea is using the dataset-extracted domain knowledge to occupy the medium-/high-frequency of images, thus narrowing the optimization space of adversarial perturbations. Specifically, we group the distribution of pretraining data and optimize a set of cluster-specific visual prompts on frequency domain. These prompts are incorporated with input images through prototype-based prompt selection during test period. Extensive evaluation shows that our method clearly boost MAE's adversarial robustness while maintaining its clean performance on ImageNet-1k classification.
Our code is available at: https://github.com/shikiw/RobustMAE.
****
HairCLIPv2: Unifying Hair Editing via Proxy Feature Blending
Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Weiming Zhang, Gang Hua, Nenghai Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23589-23599
Hair editing has made tremendous progress in recent years. Early hair editing methods use well-drawn sketches or masks to specify the editing conditions. Even though they can enable very fine-grained local control, such interaction modes are inefficient for the editing conditions that can be easily specified by language descriptions or reference images. Thanks to the recent breakthrough of cross-modal models (e.g., CLIP), HairCLIP is the first work that enables hair editing based on text descriptions or reference images. However, such text-driven and reference-driven interaction modes make HairCLIP unable to support fine-grained controls specified by sketch or mask. In this paper, we propose HairCLIPv2, aiming to support all the aforementioned interactions with one unified framework. Simultaneously, it improves upon HairCLIP with better irrelevant attributes (e.g., identity, background) preservation and unseen text descriptions support. The key idea is to convert all the hair editing tasks into hair transfer tasks, with editing conditions converted into different proxies accordingly. The editing effects are added upon the input image by blending the corresponding proxy features within the hairstyle or hair color feature spaces. Besides the unprecedented user interaction mode support, quantitative and qualitative experiments demonstrate the superiority of HairCLIPv2 in terms of editing effects, irrelevant attribute preservation and visual naturalness. Our code is available at https://github.com/wty-ustc/HairCLIPv2.
****
VLSlice: Interactive Vision-and-Language Slice Discovery
Eric Slyman, Minsuk Kahng, Stefan Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15291-15301
Recent work in vision-and-language demonstrates that large-scale pretraining can learn generalizable models that are efficiently transferable to downstream tasks. While this may improve dataset-scale aggregate metrics, analyzing performance

around hand-crafted subgroups targeting specific bias dimensions reveals system
ic undesirable behaviors. However, this subgroup analysis is frequently stalled
by annotation efforts, which require extensive time and resources to collect the
 necessary data. Prior art attempts to automatically discover subgroups to circu
mvent these constraints but typically leverages model behavior on existing task-
specific annotations and rapidly degrades on more complex inputs beyond "tabular
" data, none of which study vision-and-language models. This paper presents VLSl
ice, an interactive system enabling user-guided discovery of coherent representa
tion-level subgroups with consistent visiolinguistic behavior, denoted as vision
-and-language slices, from unlabeled image sets. We show that VLSlice enables us
ers to quickly generate diverse high-coherency slices in a user study (n=22) and
 release the tool publicly.
****
Learning to Ground Instructional Articles in Videos through Narrations
Effrosyni Mavroudi, Triantafyllos Afouras, Lorenzo Torresani; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15201-15
213
In this paper we present an approach for localizing steps of procedural activiti
es in narrated how-to videos. To deal with the scarcity of labeled data at scale
, we source the step descriptions from a language knowledge base (wikiHow) conta
ining instructional articles for a large variety of procedural tasks. Without an
y form of manual supervision, our model learns to temporally ground the steps of
 procedural articles in how-to videos by matching three modalities: frames, narr
ations, and step descriptions. Specifically, our method
 aligns steps to video by fusing information from two distinct pathways: i) dire
ct alignment of step descriptions to frames, ii) indirect alignment obtained by
composing steps-to-narrations with narrations-to-video correspondences.
 Notably, our approach performs global temporal grounding of all steps in an art
icle at once by exploiting order information, and is trained with step pseudo-la
bels which are iteratively refined and aggressively filtered.
 In order to validate our model we introduce a new benchmark -- HT-Step -- obtai
ned by manually annotating a 124-hour subset of HowTo100M with steps sourced fro
m wikiHow articles. Experiments on this benchmark as well as zero-shot evaluatio
ns on CrossTask demonstrate that our multi-modality alignment yields dramatic ga
ins over several baselines and prior works. Finally, we show that our inner modu
le for matching narration-to-video outperforms by a large margin the state of th
e art on the HTM-Align narration-video alignment benchmark.
****
DocTr: Document Transformer for Structured Information Extraction in Documents
Haofu Liao, Aruni RoyChowdhury, Weijian Li, Ankan Bansal, Yuting Zhang, Zhuowen
Tu, Ravi Kumar Satzoda, R. Manmatha, Vijay Mahadevan; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2023, pp. 19584-19594
We present a new formulation for structured information extraction (SIE) from vi
sually rich documents. We address the limitations of existing IOB tagging and gr
aph-based formulations, which are either overly reliant on the correct ordering
of input text or struggle with decoding a complex graph. Instead, motivated by a
nchor-based object detectors in computer vision, we represent an entity as an an
chor word and a bounding box, and represent entity linking as the association be
tween anchor words. This is more robust to text ordering, and maintains a compac
t graph for entity linking. The formulation motivates us to introduce 1) a Docum
ent Transformer (DocTr) that aims at detecting and associating entity bounding b
oxes in visually rich documents, and 2) a simple pre-training strategy that help
s learn entity detection in the context of language. Evaluations on three SIE be
nchmarks show the effectiveness of the proposed formulation, and the overall app
roach outperforms existing solutions.
****
The Making and Breaking of Camouflage
Hala Lamdouar, Weidi Xie, Andrew Zisserman; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 832-842
Not all camouflages are equally effective, as even a partially visible contour o

r a slight color difference can make the animal stand out and break its camoufla
ge. In this paper, we address the question of what makes a camouflage successful
, by proposing three scores for automatically assessing its effectiveness. In pa
rticular, we show that camouflage can be measured by the similarity between back
ground and foreground features and boundary visibility. We use these camouflage
scores to assess and compare all available camouflage datasets. We also incorpor
ate the proposed camouflage score into a generative model as an auxiliary loss a
nd show that effective camouflage images or videos can be synthesised in a scala
ble manner. The generated synthetic dataset is used to train a transformer-based
 model for segmenting camouflaged animals in videos. Experimentally, we demonstr
ate state-of-the-art camouflage breaking performance on the public MoCA-Mask ben
chmark.
****
Role-Aware Interaction Generation from Textual Description
Mikihiro Tanaka, Kent Fujiwara; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 15999-16009
This research tackles the problem of generating interaction between two human ac
tors corresponding to textual description. We claim that certain interactions, w
hich we call asymmetric interactions, involve a relationship between an actor an
d a receiver, whose motions significantly differ depending on the assigned role.
 However, existing studies of interaction generation attempt to learn the corres
pondence between a single label and the motions of both actors combined, overloo
king differences in individual roles. We consider a novel problem of role-aware
interaction generation, where roles can be designated before generation. We tran
slate the text of the asymmetric interactions into active and passive voice to e
nsure the textual context is consistent with each role. We propose a model that
learns to generate motions of the designated role, which together form a mutuall
y consistent interaction. As the model treats individual motions separately, it
can be pretrained to derive knowledge from single-person motion data for more ac
curate interactions. Moreover, we introduce a method inspired by Permutation Inv
ariant Training (PIT) that can automatically learn which of the two actions corr
esponds to an actor or a receiver without additional annotation. We further pres
ent cases where existing evaluation metrics fail to accurately assess the qualit
y of generated interactions, and propose a novel metric, Mutual Consistency, to
address such shortcomings. Experimental results demonstrate the efficacy of our
method, as well as the necessity of the proposed metric. Our code is available a
t https://github.com/line/Human-Interaction-Generation.
****
MAMo: Leveraging Memory and Attention for Monocular Video Depth Estimation
Rajeev Yasarla, Hong Cai, Jisoo Jeong, Yunxiao Shi, Risheek Garrepalli, Fatih Po
rikli; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2023, pp. 8754-8764
We propose MAMo, a novel memory and attention framework for monocular video dept
h estimation. MAMo can augment and improve any single-image depth estimation net
works into video depth estimation models, enabling them to take advantage of the
 temporal information to predict more accurate depth. In MAMo, we augment model
with memory which aids the depth prediction as the model streams through the vid
eo. Specifically, the memory stores learned visual and displacement tokens of th
e previous time instances. This allows the depth network to cross-reference rele
vant features from the past when predicting depth on the current frame. We intro
duce a novel scheme to continuously update the memory, optimizing it to keep tok
ens that correspond with both the past and the present visual information. We ad
opt attention-based approach to process memory features where we first learn the
 spatio-temporal relation among the resultant visual and displacement memory tok
ens using self-attention module. Further, the output features of self-attention
are aggregated with the current visual features through cross-attention. The cro
ss-attended features are finally given to a decoder to predict depth on the curr
ent frame. Through extensive experiments on several benchmarks, including KITTI,
 NYU-Depth V2, and DDAD, we show that MAMo consistently improves monocular depth
 estimation networks and sets new state-of-the-art (SOTA) accuracy. Notably, our

MAMo video depth estimation provides higher accuracy with lower latency, when comparing to SOTA cost-volume-based video depth models.
****
Continual Learning for Personalized Co-speech Gesture Generation
Chaitanya Ahuja, Pratik Joshi, Ryo Ishii, Louis-Philippe Morency; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20893-20903
Co-speech gestures are a key channel of human communication, making them important for personalized chat agents to generate. In the past, gesture generation models assumed that data for each speaker is available all at once, and in large amounts. However in practical scenarios, speaker data comes sequentially and in small amounts as the agent personalizes with more speakers, akin to a continual learning paradigm. While more recent works have shown progress in adapting to low-resource data, they catastrophically forget the gesture styles of initial speakers they were trained on. Also, prior generative continual learning works are not multimodal, making this space less studied. In this paper, we explore this new paradigm and propose C-DiffGAN: an approach that continually learns new speaker gesture styles with only a few minutes of per-speaker data, while retaining previously learnt styles. Inspired by prior continual learning works, C-DiffGAN encourages knowledge retention by 1) generating reminiscences of previous low-resource speaker data, then 2) crossmodally aligning to them to mitigate catastrophic forgetting. We quantitatively demonstrate improved performance and reduced forgetting over strong baselines through standard continual learning measures, reinforced by a qualitative user study that shows that our method produces more natural, style-preserving gestures. Code and videos can be found at https://chahuja.com/cdiffgan
****
Object as Query: Lifting Any 2D Object Detector to 3D Detection
Zitian Wang, Zehao Huang, Jiahui Fu, Naiyan Wang, Si Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3791-3800
3D object detection from multi-view images has drawn much attention over the past few years. Existing methods mainly establish 3D representations from multi-view images and adopt a dense detection head for object detection, or employ object queries distributed in 3D space to localize objects. In this paper, we design Multi-View 2D Objects guided 3D Object Detector (MV2D), which can lift any 2D object detector to multi-view 3D object detection. Since 2D detections can provide valuable priors for object existence, MV2D exploits 2D detectors to generate object queries conditioned on the rich image semantics. These dynamically generated queries help MV2D to recall objects in the field of view and show a strong capability of localizing 3D objects. For the generated queries, we design a sparse cross attention module to force them to focus on the features of specific objects, which suppresses interference from noises. The evaluation results on the nuScenes dataset demonstrate the dynamic object queries and sparse feature aggregation can promote 3D detection capability. MV2D also exhibits a state-of-the-art performance among existing methods. We hope MV2D can serve as a new baseline for future research.
****
HDG-ODE: A Hierarchical Continuous-Time Model for Human Pose Forecasting
Yucheng Xing, Xin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14700-14712
Recently, human pose estimation has attracted more and more attention due to its importance in many real applications. Although many efforts have been put on extracting 2D poses from static images, there are still some severe problems to be solved. A critical one is occlusion, which is more obvious in multi-person scenarios and makes it even more difficult to recover the corresponding 3D poses. When we consider a sequence of images, the temporal correlation among the contexts can be utilized to help us ease the problem, but most of the current works only rely on discrete-time models and estimate the joint locations of all people within a whole sparse graph. In this paper, we propose a new framework, Hierarchical Dynamic Graph Ordinary Differential Equation (HDG-ODE), to tackle the 3D pose

forecasting task from 2D skeleton representations in videos. Our framework adopts ODE, a continuous-time model, as the base to predict the 3D joint positions at any time. Considering the structural-property of the skeleton data in representing human poses and the possible irregularity caused by occlusion, we propose the use of dynamic graph convolution as the basic operator. To reduce the computational complexity introduced by the sparsity of the pose graph, our model takes a hierarchical structure where the encoding process at the observation timestamp is done in a cascade manner while the propagation between observations is conducted in parallel. The performance studies on several datasets demonstrate that our model is effective and can out-perform other methods with fewer parameters.
****
Versatile Diffusion: Text, Images and Variations All in One Diffusion Model
Xingqian Xu, Zhangyang Wang, Gong Zhang, Kai Wang, Humphrey Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7754-7765
Recent advances in diffusion models have set an impressive milestone in many generation tasks, and trending works such as DALL-E2, Imagen, and Stable Diffusion have attracted great interest. Despite the rapid landscape changes, recent new approaches focus on extensions and performance rather than capacity, thus requiring separate models for separate tasks. In this work, we expand the existing single-flow diffusion pipeline into a multi-task multimodal network, dubbed Versatile Diffusion (VD), that handles multiple flows of text-to-image, image-to-text, and variations in one unified model. The pipeline design of VD instantiates a unified multi-flow diffusion framework, consisting of sharable and swappable layer modules that enable the crossmodal generality beyond images and text. Through extensive experiments, we demonstrate that VD successfully achieves the following: a) VD outperforms the baseline approaches and handles all its base tasks with competitive quality; b) VD enables novel extensions such as disentanglement of style and semantics, dual- and multi-context blending, etc.; c) The success of our multi-flow multimodal framework over images and text may inspire further diffusion-based universal AI research. Our code and models are open-sourced at https://github.com/SHI-Labs/Versatile-Diffusion.
****
DreamTeacher: Pretraining Image Backbones with Deep Generative Models
Daiqing Li, Huan Ling, Amlan Kar, David Acuna, Seung Wook Kim, Karsten Kreis, Antonio Torralba, Sanja Fidler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16698-16708
In this work, we introduce a self-supervised feature representation learning framework DreamTeacher that utilizes generative networks for pre-training downstream image backbones. We propose to distill knowledge from a trained generative model into standard image backbones that have been well engineered for specific perception tasks. We investigate two types of knowledge distillation: 1) distilling learned generative features onto target image backbones as an alternative to pretraining these backbones on large labeled datasets such as ImageNet, and 2) distilling labels obtained from generative networks with task heads onto logits of target backbones. We perform extensive analysis on several generative models, dense prediction benchmarks, and several pre-training regimes. We empirically find that our DreamTeacher significantly outperforms existing self-supervised representation learning approaches across the board. Unsupervised ImageNet pre-training with DreamTeacher leads to significant improvements over ImageNet classification pre-training on downstream datasets, showcasing generative models, and diffusion generative models specifically, as a promising approach to representation learning on large, diverse datasets without requiring manual annotation.
****
Decomposition-Based Variational Network for Multi-Contrast MRI Super-Resolution and Reconstruction
Pengcheng Lei, Faming Fang, Guixu Zhang, Tieyong Zeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21296-21306
Multi-contrast MRI super-resolution (SR) and reconstruction methods aim to explore complementary information from the reference image to help the reconstruction

of the target image. Existing deep learning-based methods usually manually design fusion rules to aggregate the multi-contrast images, fail to model their correlations accurately and lack certain interpretations. Against these issues, we propose a multi-contrast variational network (MC-VarNet) to explicitly model the relationship of multi-contrast images. Our model is constructed based on an intuitive motivation that multi-contrast images have consistent (edges and structures) and inconsistent (contrast) information. We thus build a model to reconstruct the target image and decompose the reference image as a common component and a unique component. In the feature interaction phase, only the common component is transferred to the target image. We solve the variational model and unfold the iterative solutions into a deep network. Hence, the proposed method combines the good interpretability of model-based methods with the powerful representation ability of deep learning-based methods. Experimental results on the multi-contrast MRI reconstruction and SR demonstrate the effectiveness of the proposed model. Especially, since we explicitly model the multi-contrast images, our model is more robust to the reference images with noises and large inconsistent structures. The code is available at https://github.com/lpcccc-cv/MC-VarNet.
****
Self-supervised Monocular Underwater Depth Recovery, Image Restoration, and a Real-sea Video Dataset
Nisha Varghese, Ashish Kumar, A. N. Rajagopalan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12248-12258
Underwater (UW) depth estimation and image restoration is a challenging task due to its fundamental ill-posedness and the unavailability of real large-scale UW-paired datasets. UW depth estimation has been attempted before by utilizing either the haze information present or the geometry cue from stereo images or the adjacent frames in a video. To obtain improved estimates of depth from a single UW image, we propose a deep learning (DL) method that utilizes both haze and geometry during training. By harnessing the physical model for UW image formation in conjunction with the view-synthesis constraint on neighboring frames in monocular videos, we perform disentanglement of the input image to also get an estimate of the scene radiance. The proposed method is completely self-supervised and simultaneously outputs the depth map and the restored image in real-time (55 fps). We call this first-ever Underwater Self-supervised deep learning network for simultaneous Recovery of Depth and Image as USe-ReDI-Net. To facilitate monocular self-supervision, we collected a Dataset of Real-world Underwater Videos of Artifacts (DRUVA) in shallow sea waters. DRUVA is the first UW video dataset that contains video sequences of 20 different submerged artifacts with almost full azimuthal coverage of each artifact. Extensive experiments on our DRUVA dataset and other UW datasets establish the superiority of our proposed USe-ReDI-Net over prior art for both UW depth and image recovery. The dataset DRUVA is available at https://github.com/nishavarghese15/DRUVA
****
Geometrized Transformer for Self-Supervised Homography Estimation
Jiazhen Liu, Xirong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9556-9565
For homography estimation, we propose Geometrized Transformer (GeoFormer), a new detector-free feature matching method. Current detector-free methods, e.g. LoFTR, lack an effective mean to accurately localize small and thus computationally feasible regions for cross-attention diffusion. We resolve the challenge with an extremely simple idea: using the classical RANSAC geometry for attentive region search. Given coarse matches by LoFTR, a homography is obtained with ease. Such a homography allows us to compute cross-attention in a focused manner, where key/value sets required by Transformers can be reduced to small fix-sized regions rather than an entire image. Local features can thus be enhanced by standard Transformers. We integrate GeoFormer into the LoFTR framework. By minimizing a multi-scale cross-entropy based matching loss on auto-generated training data, the network is trained in a fully self-supervised manner. Extensive experiments are conducted on multiple real-world datasets covering natural images, heavily manipulated pictures and retinal images. The proposed method compares favorably agains

t the state-of-the-art.
****
Sat2Density: Faithful Density Learning from Satellite-Ground Image Pairs
Ming Qian, Jincheng Xiong, Gui-Song Xia, Nan Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3683-3692
This paper aims to develop an accurate 3D geometry representation of satellite images using satellite-ground image pairs. Our focus is on the challenging problem of 3D-aware ground-views synthesis from a satellite image. We draw inspiration from the density field representation used in volumetric neural rendering and propose a new approach, called Sat2Density. Our method utilizes the properties of ground-view panoramas for the sky and non-sky regions to learn faithful density fields of 3D scenes in a geometric perspective. Unlike other methods that require extra depth information during training, our Sat2Density can automatically learn accurate and faithful 3D geometry via density representation without depth supervision. This advancement significantly improves the ground-view panorama synthesis task. Additionally, our study provides a new geometric perspective to understand the relationship between satellite and ground-view images in 3D space.
****
TiDy-PSFs: Computational Imaging with Time-Averaged Dynamic Point-Spread-Functions
Sachin Shah, Sakshum Kulshrestha, Christopher A. Metzler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10657-10667
Point-spread-function (PSF) engineering is a powerful computational imaging technique wherein a custom phase mask is integrated into an optical system to encode additional information into captured images. Used in combination with deep learning, such systems now offer state-of-the-art performance at monocular depth estimation, extended depth-of-field imaging, lensless imaging, and other tasks. Inspired by recent advances in spatial light modulator (SLM) technology, this paper answers a natural question: Can one encode additional information and achieve superior performance by changing a phase mask dynamically over time? We first prove that the set of PSFs described by static phase masks is non-convex and that, as a result, time-averaged PSFs generated by dynamic phase masks are fundamentally more expressive. We then demonstrate, in simulation, that time-averaged dynamic (TiDy) phase masks can leverage this increased expressiveness to offer substantially improved monocular depth estimation and extended depth-of-field imaging performance.
****
Expressive Text-to-Image Generation with Rich Text
Songwei Ge, Taesung Park, Jun-Yan Zhu, Jia-Bin Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7545-7556
Plain text has become a prevalent interface for text-to-image synthesis. However, its limited customization options hinder users from accurately describing desired outputs. For example, plain text makes it hard to specify continuous quantities, such as the precise RGB color value or importance of each word. Furthermore, creating detailed text prompts for complex scenes is tedious for humans to write and challenging for text encoders to interpret. To address these challenges, we propose using a rich-text editor supporting formats such as font style, size, color, and footnote. We extract each word's attributes from rich text to enable local style control, explicit token reweighting, precise color rendering, and detailed region synthesis. We achieve these capabilities through a region-based diffusion process. We first obtain each word's region based on attention maps of a diffusion process using plain text. For each region, we enforce its text attributes by creating region-specific detailed prompts and applying region-specific guidance, and maintain its fidelity against plain-text generation through region-based injections. We present various examples of image generation from rich text and demonstrate that our method outperforms strong baselines with quantitative evaluations.
****
Learning Fine-Grained Features for Pixel-Wise Video Correspondences
Rui Li, Shenglong Zhou, Dong Liu; Proceedings of the IEEE/CVF International Conf

erence on Computer Vision (ICCV), 2023, pp. 9632-9641
Video analysis tasks rely heavily on identifying the pixels from different frames that correspond to the same visual target. To tackle this problem, recent studies have advocated feature learning methods that aim to learn distinctive representations to match the pixels, especially in a self-supervised fashion. Unfortunately, these methods have difficulties for tiny or even single-pixel visual targets. Pixel-wise video correspondences were traditionally related to optical flows, which however lead to deterministic correspondences and lack robustness on real-world videos. We address the problem of learning features for establishing pixel-wise correspondences. Motivated by optical flows as well as the self-supervised feature learning, we propose to use not only labeled synthetic videos but also unlabeled real-world videos for learning fine-grained representations in a holistic framework. We adopt an adversarial learning scheme to enhance the generalization ability of the learned features. Moreover, we design a coarse-to-fine framework to pursue high computational efficiency. Our experimental results on a series of correspondence-based tasks demonstrate that the proposed method outperforms state-of-the-art rivals in both accuracy and efficiency.
****
FS-DETR: Few-Shot DEtection TRansformer with Prompting and without Re-Training
Adrian Bulat, Ricardo Guerrero, Brais Martinez, Georgios Tzimiropoulos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11793-11802
This paper is on Few-Shot Object Detection (FSOD), where given a few templates (examples) depicting a novel class (not seen during training), the goal is to detect all of its occurrences within a set of images. From a practical perspective, an FSOD system must fulfil the following desiderata: (a) it must be used as is, without requiring any fine-tuning at test time, (b) it must be able to process an arbitrary number of novel objects concurrently while supporting an arbitrary number of examples from each class and (c) it must achieve accuracy comparable to a closed system. Towards satisfying (a)-(c), in this work, we make the following contributions: We introduce, for the first time, a simple, yet powerful, few-shot detection transformer (FS-DETR) based on visual prompting that can address both desiderata (a) and (b). Our system builds upon the DETR framework, extending it based on two key ideas: (1) feed the provided visual templates of the novel classes as visual prompts during test time, and (2) "stamp" these prompts with pseudo-class embeddings (akin to soft prompting), which are then predicted at the output of the decoder. Importantly, we show that our system is not only more flexible than existing methods, but also, it makes a step towards satisfying desideratum (c). Specifically, it is significantly more accurate than all methods that do not require fine-tuning and even matches and outperforms the current state-of-the-art fine-tuning based methods on the most well-established benchmarks (PASCAL VOC & MSCOCO).
****
Semi-supervised Speech-driven 3D Facial Animation via Cross-modal Encoding
Peiji Yang, Huawei Wei, Yicheng Zhong, Zhisheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21032-21041
Existing Speech-driven 3D facial animation methods typically follow the supervised paradigm, involving regression from speech to 3D facial animation. This paradigm faces two major challenges: the high cost of supervision acquisition, and the ambiguity in mapping between speech and lip movements. To address these challenges, this study proposes a novel cross-modal semi-supervised framework, comprising a Speech-to-Image Transcoder and a Face-to-Geometry Regressor. The former jointly learns a common representation space from speech and image domains, enabling the transformation of speech into semantically-consistent facial images. The latter is responsible for reconstructing 3D facial meshes from the transformed images. Both modules require minimal effort to acquire the necessary training data, thereby obviating the dependence on costly supervised data. Furthermore, the joint learning scheme enables the fusion of intricate visual features into speech encoding, thereby facilitating the transformation of subtle speech variations into nuanced lip movements, ultimately enhancing the fidelity of 3D face reconst

ructions. Consequently, the ambiguity of the direct mapping of speech-to-animati on is significantly reduced, leading to coherent and high-fidelity generation of lip motion. Extensive experiments demonstrate that our approach produces compet itive results compared to supervised methods.
****
Learning to Learn: How to Continuously Teach Humans and Machines
Parantak Singh, You Li, Ankur Sikarwar, Stan Weixian Lei, Difei Gao, Morgan B. T albot, Ying Sun, Mike Zheng Shou, Gabriel Kreiman, Mengmi Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11708 -11719
Curriculum design is a fundamental component of education. For example, when we learn mathematics at school, we build upon our knowledge of addition to learn mu ltiplication. These and other concepts must be mastered before our first algebra lesson, which also reinforces our addition and multiplication skills. Designing a curriculum for teaching either a human or a machine shares the underlying goa l of maximizing knowledge transfer from earlier to later tasks, while also minim izing forgetting of learned tasks. Prior research on curriculum design for image classification focuses on the ordering of training examples during a single off line task. Here, we investigate the effect of the order in which multiple distin ct tasks are learned in a sequence. We focus on the online class-incremental con tinual learning setting, where algorithms or humans must learn image classes one at a time during a single pass through a dataset. We find that curriculum consi stently influences learning outcomes for humans and for multiple continual machi ne learning algorithms across several benchmark datasets. We introduce a novel-o bject recognition dataset for human curriculum learning experiments and observe that curricula that are effective for humans are highly correlated with those th at are effective for machines. As an initial step towards automated curriculum d esign for online class-incremental learning, we propose a novel algorithm, dubbe d Curriculum Designer (CD), that designs and ranks curricula based on inter-clas s feature similarities. We find significant overlap between curricula that are e mpirically highly effective and those that are highly ranked by our CD. Our stud y establishes a framework for further research on teaching humans and machines t o learn continuously using optimized curricula.
****
Text-Driven Generative Domain Adaptation with Spectral Consistency Regularizatio n
Zhenhuan Liu, Liang Li, Jiayu Xiao, Zheng-Jun Zha, Qingming Huang; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 701 9-7029
Combined with the generative prior of pre-trained models and the flexibility of text, text-driven generative domain adaptation can generate images from a wide r ange of target domains. However, current methods still suffer from overfitting a nd the mode collapse problem. In this paper, we analyze the mode collapse from t he geometric point of view and reveal its relationship to the Hessian matrix of generator. To alleviate it, we propose the spectral consistency regularization t o preserve the diversity of source domain without restricting the semantic adapt ation to target domain. We also design granularity adaptive regularization to fl exibly control the balance between diversity and stylization for target model. W e conduct experiments for broad target domains compared with state-of-the-art me thods and extensive ablation studies. The experiments demonstrate the effectiven ess of our method to preserve the diversity of source domain and generate high f idelity target images.
****
A 5-Point Minimal Solver for Event Camera Relative Motion Estimation
Ling Gao, Hang Su, Daniel Gehrig, Marco Cannici, Davide Scaramuzza, Laurent Knei p; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8049-8059
Event-based cameras are ideal for line-based motion estimation, since they predo minantly respond to edges in the scene. However, accurately determining the came ra displacement based on events continues to be an open problem. This is because

line feature extraction and dynamics estimation are tightly coupled when using event cameras, and no precise model is currently available for describing the complex structures generated by lines in the space-time volume of events. We solve this problem by deriving the correct non-linear parametrization of such manifolds, which we term eventails, and demonstrate its application to event-based linear motion estimation, with known rotation from an Inertial Measurement Unit. Using this parametrization, we introduce a novel minimal 5-point solver that jointly estimates line parameters and linear camera velocity projections, which can be fused into a single, averaged linear velocity when considering multiple lines. We demonstrate on both synthetic and real data that our solver generates more stable relative motion estimates than other methods while capturing more inliers than clustering based on spatio-temporal planes. In particular, our method consistently achieves a 100% success rate in estimating linear velocity where existing closed-form solvers only achieve between 23% and 70%. The proposed eventails contribute to a better understanding of spatio-temporal event-generated geometries and we thus believe it will become a core building block of future event-based motion estimation algorithms.
****
TM2D: Bimodality Driven 3D Dance Generation via Music-Text Integration
Kehong Gong, Dongze Lian, Heng Chang, Chuan Guo, Zihang Jiang, Xinxin Zuo, Michael Bi Mi, Xinchao Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9942-9952
We propose a novel task for generating 3D dance movements that simultaneously incorporate both text and music modalities. Unlike existing works that generate dance movements using a single modality such as music, our goal is to produce richer dance movements guided by the instructive information provided by the text. However, the lack of paired motion data with both music and text modalities limits the ability to generate dance movements that integrate both. To alleviate this challenge, we propose to utilize a 3D human motion VQ-VAE to project the motions of the two datasets into a latent space consisting of quantized vectors, which effectively mix the motion tokens from the two datasets with different distributions for training. Additionally, we propose a cross-modal transformer to integrate text instructions into motion generation architecture for generating 3D dance movements without degrading the performance of music-conditioned dance generation. To better evaluate the quality of the generated motion, we introduce two novel metrics, namely Motion Prediction Distance (MPD) and Freezing Score (FS), to measure the coherence and freezing percentage of the generated motion. Extensive experiments show that our approach can generate realistic and coherent dance movements conditioned on both text and music while maintaining comparable performance with the two single modalities. Code is available at https://garfield-kh.github.io/TM2D/.
****
Bootstrap Motion Forecasting With Self-Consistent Constraints
Maosheng Ye, Jiamiao Xu, Xunnong Xu, Tengfei Wang, Tongyi Cao, Qifeng Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8504-8514
We present a novel framework to bootstrap Motion forecasting with Self-consistent Constraints (MISC). The motion forecasting task aims at predicting future trajectories of vehicles by incorporating spatial and temporal information from the past. A key design of MISC is the proposed Dual Consistency Constraints that regularize the predicted trajectories under spatial and temporal perturbation during training. Also, to model the multi-modality in motion forecasting, we design a novel self-ensembling scheme to obtain accurate teacher targets to enforce the self-constraints with multi-modality supervision. With explicit constraints from multiple teacher targets, we observe a clear improvement in the prediction performance. Extensive experiments on the Argoverse motion forecasting benchmark and Waymo Open Motion dataset show that MISC significantly outperforms the state-of-the-art methods. As the proposed strategies are general and can be easily incorporated into other motion forecasting approaches, we also demonstrate that our proposed scheme consistently improves the prediction performance of several exist

ing methods.
****
CDAC: Cross-domain Attention Consistency in Transformer for Domain Adaptive Semantic Segmentation
Kaihong Wang, Donghyun Kim, Rogerio Feris, Margrit Betke; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11519-11529
While transformers have greatly boosted performance in semantic segmentation, domain adaptive transformers are not yet well explored. We identify that the domain gap can cause discrepancies in self-attention. Due to this gap, the transformer attends to spurious regions or pixels, which deteriorates accuracy on the target domain. We propose Cross-Domain Attention Consistency (CDAC), to perform adaptation on attention maps using cross-domain attention layers that share features between source and target domains. Specifically, we impose consistency between predictions from cross-domain attention and self-attention modules to encourage similar distributions across domains in both the attention and output of the model, i.e., attention-level and output-level alignment. We also enforce consistency in attention maps between different augmented views to further strengthen the attention-based alignment. Combining these two components, CDAC mitigates the discrepancy in attention maps across domains and further boosts the performance of the transformer under unsupervised domain adaptation settings. Our method is evaluated on various widely used benchmarks and outperforms the state-of-the-art baselines, including GTAV-to-Cityscapes by 1.3 and 1.5 percent point (pp) and Synthia-to-Cityscapes by 0.6 pp and 2.9 pp when combining with two competitive Transformer-based backbones, respectively. Our code will be publicly available at https://github.com/wangkaihong/CDAC.
****
WaveNeRF: Wavelet-based Generalizable Neural Radiance Fields
Muyu Xu, Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Xiaoqin Zhang, Christian Theobalt, Ling Shao, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18195-18204
Neural Radiance Field (NeRF) has shown impressive performance in novel view synthesis via implicit scene representation. However, it usually suffers from poor scalability as requiring densely sampled images for each new scene. Several studies have attempted to mitigate this problem by integrating Multi-View Stereo (MVS) technique into NeRF while they still entail a cumbersome fine-tuning process for new scenes. Notably, the rendering quality will drop severely without this fine-tuning process and the errors mainly appear around the high-frequency features. In the light of this observation, we design WaveNeRF, which integrates wavelet frequency decomposition into MVS and NeRF to achieve generalizable yet high-quality synthesis without any per-scene optimization. To preserve high-frequency information when generating 3D feature volumes, WaveNeRF builds Multi-View Stereo in the Wavelet domain by integrating the discrete wavelet transform into the classical cascade MVS, which disentangles high-frequency information explicitly. With that, disentangled frequency features can be injected into classic NeRF via a novel hybrid neural renderer to yield faithful high-frequency details, and an intuitive frequency-guided sampling strategy can be designed to suppress artifacts around high-frequency regions. Extensive experiments over three widely studied benchmarks show that WaveNeRF achieves superior generalizable radiance field modeling when only given three images as input.
****
LoCUS: Learning Multiscale 3D-consistent Features from Posed Images
Dominik A. Kloepfer, Dylan Campbell, João F. Henriques; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16634-16644
An important challenge for autonomous agents such as robots is to maintain a spatially and temporally consistent model of the world. It must be maintained through occlusions, previously-unseen views, and long time horizons (e.g., loop closure and re-identification). It is still an open question how to train such a versatile neural representation without supervision.
We start from the idea that the training objective can be framed as a patch retrieval problem: given an image patch in one view of a scene, we would like to re

trieve (with high precision and recall) all patches in other views that map to the same real-world location. One drawback is that this objective does not promote reusability of features: by being unique to a scene (achieving perfect precision/recall), a representation will not be useful in the context of other scenes. We find that it is possible to balance retrieval and reusability by constructing the retrieval set carefully, leaving out patches that map to far-away locations. Similarly, we can easily regulate the scale of the learned features (e.g., points, objects, or rooms) by adjusting the spatial tolerance for considering a retrieval to be positive. We optimize for (smooth) Average Precision (AP), in a single unified ranking-based objective. This objective also doubles as a criterion for choosing landmarks or keypoints, as patches with high AP.

We show results creating sparse, multi-scale, semantic spatial maps composed of highly identifiable landmarks, with applications in landmark retrieval, localization, semantic segmentation and instance segmentation.

****

Neural Reconstruction of Relightable Human Model from Monocular Video

Wenzhang Sun, Yunlong Che, Han Huang, Yandong Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 397-407

Creating relightable and animatable human characters from monocular video at a low cost is a critical task for digital human modeling and virtual reality applications. This task is complex due to intricate articulation motion, a wide range of ambient lighting conditions, and pose-dependent clothing deformations. In this paper, we introduce a novel self-supervised framework that takes a monocular video of a moving human as input and generates a 3D neural representation capable of being rendered with novel poses under arbitrary lighting conditions. Our framework decomposes dynamic humans under varying illumination into neural fields in canonical space, taking into account geometry and spatially varying BRDF material properties. Additionally, we introduce pose-driven deformation fields, enabling bidirectional mapping between canonical space and observation. Leveraging the proposed appearance decomposition and deformation fields, our framework learns in a self-supervised manner. Ultimately, based on pose-driven deformation, recovered appearance, and physically-based rendering, the reconstructed human figure becomes relightable and can be explicitly driven by novel poses. We demonstrate significant performance improvements over previous works and provide compelling examples of relighting from monocular videos of moving humans in challenging, uncontrolled capture scenarios.

****

FB-BEV: BEV Representation from Forward-Backward View Transformations

Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, Jose M. Alvarez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6919-6928

View Transformation Module (VTM), where transformations happen between multi-view image features and Bird-Eye-View (BEV) representation, is a crucial step in camera-based BEV perception systems. Currently, the two most prominent VTM paradigms are forward projection and backward projection. Forward projection, represented by Lift-Splat-Shoot, leads to sparsely projected BEV features without post-processing. Backward projection, with BEVFormer being an example, tends to generate false-positive BEV features from incorrect projections due to the lack of utilization on depth.

To address the above limitations, we propose a novel forward-backward view transformation module. Our approach compensates for the deficiencies in both existing methods, allowing them to enhance each other to obtain higher quality BEV representations mutually. We instantiate the proposed module with FB-BEV, which achieves a new state-of-the-art result of 62.4% NDS on the nuScenes test set. Code and models are available at https://github.com/NVlabs/FB-BEV

****

BoxSnake: Polygonal Instance Segmentation with Box Supervision

Rui Yang, Lin Song, Yixiao Ge, Xiu Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 766-776

Box-supervised instance segmentation has gained much attention as it requires on

ly simple box annotations instead of costly mask or polygon annotations. However, existing box-supervised instance segmentation models mainly focus on mask-based frameworks. We propose a new end-to-end training technique, termed BoxSnake, to achieve effective polygonal instance segmentation using only box annotations for the first time. Our method consists of two loss functions: (1) a point-based unary loss that constrains the bounding box of predicted polygons to achieve coarse-grained segmentation; and (2) a distance-aware pairwise loss that encourages the predicted polygons to fit the object boundaries.
Compared with the mask-based weakly-supervised methods, BoxSnake further reduces the performance gap between the predicted segmentation and the bounding box, and shows significant superiority on the Cityscapes dataset.
****
Confidence-based Visual Dispersal for Few-shot Unsupervised Domain Adaptation
Yizhe Xiong, Hui Chen, Zijia Lin, Sicheng Zhao, Guiguang Ding; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11621-11631
Unsupervised domain adaptation aims to transfer knowledge from a fully-labeled source domain to an unlabeled target domain. However, in real-world scenarios, providing abundant labeled data even in the source domain can be infeasible due to the difficulty and high expense of annotation. To address this issue, recent works consider the Few-shot Unsupervised Domain Adaptation (FUDA) where only a few source samples are labeled, and conduct knowledge transfer via self-supervised learning methods. Yet existing methods generally overlook that the sparse label setting hinders learning reliable source knowledge for transfer. Additionally, the learning difficulty difference in target samples is different but ignored, leaving hard target samples poorly classified. To tackle both deficiencies, in this paper, we propose a novel Confidence-based Visual Dispersal Transfer learning method (C-VisDiT) for FUDA. Specifically, C-VisDiT consists of a cross-domain visual dispersal strategy that transfers only high-confidence source knowledge for model adaptation and an intra-domain visual dispersal strategy that guides the learning of hard target samples with easy ones. We conduct extensive experiments on Office-31, Office-Home, VisDA-C, and DomainNet benchmark datasets and the results demonstrate that the proposed C-VisDiT significantly outperforms state-of-the-art FUDA methods. Our code is available at https://github.com/Bostoncake/C-VisDiT.
****
Event-Guided Procedure Planning from Instructional Videos with Text Supervision
An-Lan Wang, Kun-Yu Lin, Jia-Run Du, Jingke Meng, Wei-Shi Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13565-13575
In this work, we focus on the task of procedure planning from instructional videos with text supervision, where a model aims to predict an action sequence to transform the initial visual state into the goal visual state. A critical challenge of this task is the large semantic gap between observed visual states and unobserved intermediate actions, which is ignored by previous works. Specifically, this semantic gap refers to that the contents in the observed visual states are semantically different from the elements of some action text labels in a procedure. To bridge this semantic gap, we propose a novel event-guided paradigm, which first infers events from the observed states and then plans out actions based on both the states and predicted events. Our inspiration comes from that planning a procedure from an instructional video is to complete a specific event and a specific event usually involves specific actions. Based on the proposed paradigm, we contribute an Event-guided Prompting-based Procedure Planning (E3P) model, which encodes event information into the sequential modeling process to support procedure planning. To further consider the strong action associations within each event, our E3P adopts a mask-and-predict approach for relation mining, incorporating a probabilistic masking scheme for regularization. Extensive experiments on three datasets demonstrate the effectiveness of our proposed model.
****
Foreground Object Search by Distilling Composite Image Feature

Bo Zhang, Jiacheng Sui, Li Niu; Proceedings of the IEEE/CVF International Confer ence on Computer Vision (ICCV), 2023, pp. 22986-22995
Foreground object search (FOS) aims to find compatible foreground objects for a given background image, producing realistic composite image. We observe that com petitive retrieval performance could be achieved by using a discriminator to pre dict the compatibility of composite image, but this approach has unaffordable ti me cost. To this end, we propose a novel FOS method via distilling composite fea ture (DiscoFOS). Specifically, the abovementioned discriminator serves as teache r network. The student network employs two encoders to extract foreground featur e and background feature. Their interaction output is enforced to match the comp osite image feature from the teacher network. Additionally, previous works did n ot release their datasets, so we contribute two datasets for FOS task: S-FOSD da taset with synthetic composite images and R-FOSD dataset with real composite ima ges. Extensive experiments on our two datasets demonstrate the superiority of th e proposed method over previous approaches. The dataset and code are available a t https://github.com/bcmi/Foreground-Object-Search-Dataset-FOSD.
****
Multimodal Motion Conditioned Diffusion Model for Skeleton-based Video Anomaly D etection
Alessandro Flaborea, Luca Collorone, Guido Maria D'Amely di Melendugno, Stefano D'Arrigo, Bardh Prenkaj, Fabio Galasso; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2023, pp. 10318-10329
Anomalies are rare and anomaly detection is often therefore framed as One-Class Classification (OCC), i.e. trained solely on normalcy. Leading OCC techniques co nstrain the latent representations of normal motions to limited volumes and dete ct as abnormal anything outside, which accounts satisfactorily for the openset'n ess of anomalies. But normalcy shares the same openset'ness property, since huma ns can perform the same action in several ways, which the leading techniques neg lect.
We propose a novel generative model for video anomaly detection (VAD), which as sumes that both normality and abnormality are multimodal. We consider skeletal r epresentations and leverage state-of-the-art diffusion probabilistic models to g enerate multimodal future human poses. We contribute a novel conditioning on the past motion of people and exploit the improved mode coverage capabilities of di ffusion processes to generate different-but-plausible future motions. Upon the s tatistical aggregation of future modes, an anomaly is detected when the generate d set of motions is not pertinent to the actual future. We validate our model on 4 established benchmarks: UBnormal, HR-UBnormal, HR-STC, and HR-Avenue, with ex tensive experiments surpassing state-of-the-art results.
****
ClimateNeRF: Extreme Weather Synthesis in Neural Radiance Field
Yuan Li, Zhi-Hao Lin, David Forsyth, Jia-Bin Huang, Shenlong Wang; Proceedings o f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 322 7-3238
Physical simulations produce excellent predictions of weather effects. Neural ra diance fields produce SOTA scene models. We describe a novel NeRF-editing proced ure that can fuse physical simulations with NeRF models of scenes, producing rea listic movies of physical phenomena in those scenes. Our application -- Climate NeRF -- allows people to visualize what climate change outcomes will do to them.

ClimateNeRF allows us to render realistic weather effects, including smog, snow , and flood. Results can be controlled with physically meaningful variables like water level. Qualitative and quantitative studies show that our simulated resul ts are significantly more realistic than those from SOTA 2D image editing and SO TA 3D NeRF stylization.
****
CDFSL-V: Cross-Domain Few-Shot Learning for Videos
Sarinda Samarasinghe, Mamshad Nayeem Rizve, Navid Kardan, Mubarak Shah; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp

. 11643-11652
Few-shot video action recognition is an effective approach to recognizing new ca tegories with only a few labeled examples, thereby reducing the challenges assoc iated with collecting and annotating large-scale video datasets. Existing method s in video action recognition rely on large labeled datasets from the same domai n. However, this setup is not realistic as novel categories may come from differ ent data domains that may have different spatial and temporal characteristics. T his dissimilarity between the source and target domains can pose a significant c hallenge, rendering traditional few-shot action recognition techniques ineffecti ve. To address this issue, in this work, we propose a novel cross-domain few-sho t video action recognition method that leverages self-supervised learning and cu rriculum learning to balance the information from the source and target domains. To be particular, our method employs a masked autoencoder-based self-supervised training objective to learn from both source and target data in a self-supervis ed manner. Then a progressive curriculum balances learning the discriminative in formation from the source dataset with the generic information learned from the target domain. Initially, our curriculum utilizes supervised learning to learn c lass discriminative features from the source data. As the training progresses, w e transition to learning target-domain-specific features. We propose a progressi ve curriculum to encourage the emergence of rich features in the target domain b ased on class discriminative supervised features in
the source domain. We evaluate our method on several challenging benchmark data sets and demonstrate that our approach outperforms existing cross-domain few-sho t learning techniques. Our code is available at https://github.com/Sarinda251/CD FSL-V
****
Generalized Few-Shot Point Cloud Segmentation via Geometric Words
Yating Xu, Conghui Hu, Na Zhao, Gim Hee Lee; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2023, pp. 21506-21515
Existing fully-supervised point cloud segmentation methods suffer in the dynamic testing environment with emerging new classes. Few-shot point cloud segmentatio n algorithms address this problem by learning to adapt to new classes at the sac rifice of segmentation accuracy for the base classes, which severely impedes its practicality. This largely motivates us to present the first attempt at a more practical paradigm of generalized few-shot point cloud segmentation, which requi res the model to generalize to new categories with only a few support point clou ds and simultaneously retain the capability to segment base classes. We propose the geometric words to represent geometric components shared between the base an d novel classes, and incorporate them into a novel geometric-aware semantic repr esentation to facilitate better generalization to the new classes without forget ting the old ones. Moreover, we introduce geometric prototypes to guide the segm entation with geometric prior knowledge. Extensive experiments on S3DIS and Scan Net consistently illustrate the superior performance of our method over baseline methods. Our code is available at: https://github.com/Pixie8888/GFS-3DSeg_GWs.
****
Monte Carlo Linear Clustering with Single-Point Supervision is Enough for Infrar ed Small Target Detection
Boyang Li, Yingqian Wang, Longguang Wang, Fei Zhang, Ting Liu, Zaiping Lin, Wei An, Yulan Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1009-1019
Single-frame infrared small target (SIRST) detection aims at separating small ta rgets from clutter backgrounds on infrared images. Recently, deep learning based methods have achieved promising performance on SIRST detection, but at the cost of a large amount of training data with expensive pixel-level annotations. To r educe the annotation burden, we propose the first method to achieve SIRST detect ion with single-point supervision. The core idea of this work is to recover the per-pixel mask of each target from the given single point label by using cluster ing approaches, which looks simple but is indeed challenging since targets are a lways insalient and accompanied with background clutters. To handle this issue, we introduce randomness to the clustering process by adding noise to the input i

mages, and then obtain much more reliable pseudo masks by averaging the clustered results. Thanks to this "Monte Carlo" clustering approach, our method can accurately recover pseudo masks and thus turn arbitrary fully supervised SIRST detection networks into weakly supervised ones with only single point annotation. Experiments on four datasets demonstrate that our method can be applied to existing SIRST detection networks to achieve comparable performance with their fully-supervised counterparts, which reveals that single-point supervision is strong enough for SIRST detection.
****
Practical Membership Inference Attacks Against Large-Scale Multi-Modal Models: A Pilot Study
Myeongseob Ko, Ming Jin, Chenguang Wang, Ruoxi Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4871-4881
Membership inference attacks (MIAs) aim to infer whether a data point has been used to train a machine learning model. These attacks can be employed to identify potential privacy vulnerabilities and detect unauthorized use of personal data. While MIAs have been traditionally studied for simple classification models, recent advancements in multi-modal pre-training, such as CLIP, have demonstrated remarkable zero-shot performance across a range of computer vision tasks. However, the sheer scale of data and models presents significant computational challenges for performing the attacks.

This paper takes a first step towards developing practical MIAs against large-scale multi-modal models. We introduce a simple baseline strategy by thresholding the cosine similarity between text and image features of a target point, and propose further enhancing the baseline by aggregating cosine similarity across transformations of the target. We also present a new weakly supervised attack method that leverages ground-truth non-members (e.g., obtained by using the publication date of a target model and the timestamps of the open data) to further enhance the attack. Our evaluation shows that CLIP models are susceptible to our attack strategies, with our simple baseline achieving over 75% membership identification accuracy. Furthermore, our enhanced attacks outperform the baseline across multiple models and datasets, with the weakly supervised attack demonstrating an average-case performance improvement of 17% and being at least 7X more effective at low false-positive rates. These findings highlight the importance of protecting the privacy of multi-modal foundational models, which were previously assumed to be less susceptible to MIAs due to less overfitting. The reach of the results presents unique challenges and insights for the broader community to address multi-modal privacy concerns.
****
TCOVIS: Temporally Consistent Online Video Instance Segmentation
Junlong Li, Bingyao Yu, Yongming Rao, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1097-1107
In recent years, significant progress has been made in video instance segmentation (VIS), with many offline and online methods achieving state-of-the-art performance. While offline methods have the advantage of producing temporally consistent predictions, they are not suitable for real-time scenarios. Conversely, online methods are more practical, but maintaining temporal consistency remains a challenging task. In this paper, we propose a novel online method for video instance segmentation, called TCOVIS, which fully exploits the temporal information in a video clip. The core of our method consists of a global instance assignment strategy and a spatio-temporal enhancement module, which improve the temporal consistency of the features from two aspects. Specifically, we perform global optimal matching between the predictions and ground truth across the whole video clip, and supervise the model with the global optimal objective. We also capture the spatial feature and aggregate it with the semantic feature between frames, thus realizing the spatio-temporal enhancement. We evaluate our method on four widely adopted VIS benchmarks, namely YouTube-VIS 2019/2021/2022 and OVIS, and achieve state-of-the-art performance on all benchmarks without bells-and-whistles. For instance, on YouTube-VIS 2021, TCOVIS achieves 49.5 AP and 61.3 AP with ResNet-5

0 and Swin-L backbones, respectively.
****
Towards Viewpoint Robustness in Bird's Eye View Segmentation
Tzofi Klinghoffer, Jonah Philion, Wenzheng Chen, Or Litany, Zan Gojcic, Jungseock Joo, Ramesh Raskar, Sanja Fidler, Jose M. Alvarez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8515-8524
Autonomous vehicles (AV) require that neural networks used for perception be robust to different viewpoints if they are to be deployed across many types of vehicles without the repeated cost of data collection and labeling for each. AV companies typically focus on collecting data from diverse scenarios and locations, but not camera rig configurations, due to cost. As a result, only a small number of rig variations exist across most fleets. In this paper, we study how AV perception models are affected by changes in camera viewpoint and propose a way to scale them across vehicle types without repeated data collection and labeling. Using bird's eye view (BEV) segmentation as a motivating task, we find through extensive experiments that existing perception models are surprisingly sensitive to changes in camera viewpoint. When trained with data from one camera rig, small changes to pitch, yaw, depth, or height of the camera at inference time lead to large drops in performance. We introduce a technique for novel view synthesis and use it to transform collected data to the viewpoint of target rigs, allowing us to train BEV segmentation models for diverse target rigs without any additional data collection or labeling cost. To analyze the impact of viewpoint changes, we leverage synthetic data to mitigate other gaps (content, ISP, etc). Our approach is then trained on real data and evaluated on synthetic data, enabling evaluation on diverse target rigs. We release all data for use in future work. Our method is able to recover an average of 14.7% of the IoU that is otherwise lost when deploying to new rigs.
****
Long-Term Photometric Consistent Novel View Synthesis with Diffusion Models
Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, Marcus A. Brubaker; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7094-7104
Novel view synthesis from a single input image is a challenging task, where the goal is to generate a new view of a scene from a desired camera pose that may be separated by a large motion. The highly uncertain nature of this synthesis task due to unobserved elements within the scene (i.e. occlusion) and outside the field-of-view makes the use of generative models appealing to capture the variety of possible outputs. In this paper, we propose a novel generative model capable of producing a sequence of photorealistic images consistent with a specified camera trajectory, and a single starting image. Our approach is centred on an autoregressive conditional diffusion-based model capable of interpolating visible scene elements, and extrapolating unobserved regions in a view, in a geometrically consistent manner. Conditioning is limited to an image capturing a single camera view and the (relative) pose of the new camera view. To measure the consistency over a sequence of generated views, we introduce a new metric, the thresholded symmetric epipolar distance (TSED), to measure the number of consistent frame pairs in a sequence. While previous methods have been shown to produce high quality images and consistent semantics across pairs of views, we show empirically with our metric that they are often inconsistent with the desired camera poses. In contrast, we demonstrate that our method produces both photorealistic and view-consistent imagery. Additional material is available on our project page: https://yorkucvil.github.io/Photoconsistent-NVS/.
****
What Can a Cook in Italy Teach a Mechanic in India? Action Recognition Generalisation Over Scenarios and Locations
Chiara Plizzari, Toby Perrett, Barbara Caputo, Dima Damen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13656-13666
We propose and address a new generalisation problem: can a model trained for action recognition successfully classify actions when they are performed within a previously unseen scenario and in a previously unseen location? To answer this qu

estion, we introduce the Action Recognition Generalisation Over scenarios and lo cations dataset ARGO1M, which contains 1.1M video clips from the large-scale Ego 4D dataset, across 10 scenarios and 13 locations. We demonstrate recognition mod els struggle to generalise over 10 proposed test splits, each of an unseen scena rio in an unseen location. We thus propose CIR, a method to represent each video as a Cross-Instance Reconstruction of videos from other domains. Reconstruction s are paired with text narrations to guide the learning of a domain generalisabl e representation. We provide extensive analysis and ablations on ARGO1M that sho w CIR outperforms prior domain generalisation works on all test splits.
****
EMDB: The Electromagnetic Database of Global 3D Human Pose and Shape in the Wild
Manuel Kaufmann, Jie Song, Chen Guo, Kaiyue Shen, Tianjian Jiang, Chengcheng Tan g, Juan José Zárate, Otmar Hilliges; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2023, pp. 14632-14643
We present EMDB, the Electromagnetic Database of Global 3D Human Pose and Shape in the Wild. EMDB is a novel dataset that contains high-quality 3D SMPL pose and shape parameters with global body and camera trajectories for in-the-wild video s. We use body-worn, wireless electromagnetic (EM) sensors and a hand-held iPhon e to record a total of 58 minutes of motion data, distributed over 81 indoor and outdoor sequences and 10 participants. Together with accurate body poses and sh apes, we also provide global camera poses and body root trajectories. To constru ct EMDB, we propose a multi-stage optimization procedure, which first fits SMPL to the 6-DoF EM measurements and then refines the poses via image observations. To achieve high-quality results, we leverage a neural implicit avatar model to r econstruct detailed human surface geometry and appearance, which allows for impr oved alignment and smoothness via a dense pixel-level objective. Our evaluations , conducted with a multi-view volumetric capture system, indicate that EMDB has an expected accuracy of 2.3 cm positional and 10.6 degrees angular error, surpas sing the accuracy of previous in-the-wild datasets. We evaluate existing state-o f-the-art monocular RGB methods for camera-relative and global pose estimation o n EMDB. EMDB is publicly available under https://ait.ethz.ch/emdb.
****
STEERER: Resolving Scale Variations for Counting and Localization via Selective Inheritance Learning
Tao Han, Lei Bai, Lingbo Liu, Wanli Ouyang; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2023, pp. 21848-21859
Scale variation is a deep-rooted problem in object counting, which has not been effectively addressed by existing scale-aware algorithms. An important factor is that they typically involve cooperative learning across multi-resolutions, whic h could be suboptimal for learning the most discriminative features from each sc ale. In this paper, we propose a novel method termed STEERER (SelecTivE inhERita nce lEaRning) that addresses the issue of scale variations in object counting. S TEERER selects the most suitable scale for patch objects to boost feature extrac tion and only inherits discriminative features from lower to higher resolution p rogressively. The main insights of STEERER are a dedicated Feature Selection and Inheritance Adaptor (FSIA), which selectively forwards scale-customized feature s at each scale, and a Masked Selection and Inheritance Loss (MSIL) that helps t o achieve high-quality density maps across all scales. Our experimental results on nine datasets with counting and localization tasks demonstrate the unpreceden ted scale generalization ability of STEERER. Code is available at https://github .com/taohan10200/STEERER.
****
Benchmarking Algorithmic Bias in Face Recognition: An Experimental Approach Usin g Synthetic Faces and Human Evaluation
Hao Liang, Pietro Perona, Guha Balakrishnan; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2023, pp. 4977-4987
We propose an experimental method for measuring bias in face recognition systems . Existing methods to measure bias depend on benchmark datasets that are collect ed in the wild and annotated for protected (e.g., race, gender) and non-protecte d (e.g., pose, lighting) attributes. Such observational datasets only permit cor

relational conclusions, e.g., "Algorithm A's accuracy is different on female and male faces in dataset X.". By contrast, experimental methods manipulate attributes individually and thus permit causal conclusions, e.g., "Algorithm A's accuracy is affected by gender and skin color." Our method is based on generating synthetic faces using a neural face generator, where each attribute of interest is modified independently while leaving all other attributes constant. Human observers crucially provide the ground truth on perceptual identity similarity between synthetic image pairs. We validate our method quantitatively by evaluating race and gender biases of three research-grade face recognition models. Our synthetic pipeline reveals that for these algorithms, accuracy is lower for Black and East Asian population subgroups. Our method can also quantify how perceptual changes in attributes affect face identity distances reported by these models. Our large synthetic dataset, consisting of 48,000 synthetic face image pairs (10,200 unique synthetic faces) and 555,000 human annotations (individual attributes and pairwise identity comparisons) is available to researchers in this important area.
****

Spatial-Aware Token for Weakly Supervised Object Localization

Pingyu Wu, Wei Zhai, Yang Cao, Jiebo Luo, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1844-1854

Weakly supervised object localization (WSOL) is a challenging task aiming to localize objects with only image-level supervision. Recent works apply visual transformer to WSOL and achieve significant success by exploiting the long-range feature dependency in self-attention mechanism. However, existing transformer-based methods synthesize the classification feature maps as the localization map, which leads to optimization conflicts between classification and localization tasks. To address this problem, we propose to learn a task-specific spatial-aware token (SAT) to condition localization in a weakly supervised manner. Specifically, a spatial token is first introduced in the input space to aggregate representations for localization task. Then a spatial aware attention module is constructed, which allows spatial token to generate foreground probabilities of different patches by querying and to extract localization knowledge from the classification task. Besides, for the problem of sparse and unbalanced pixel-level supervision obtained from the image-level label, two spatial constraints, including batch area loss and normalization loss, are designed to compensate and enhance this supervision. Experiments show that the proposed SAT achieves state-of-the-art performance on both CUB-200 and ImageNet, with 98.45% and 73.13% GT-known Loc, respectively. Even under the extreme setting of using only 1 image per class from ImageNet for training, SAT already exceeds the SOTA method by 2.1% GT-known Loc. Code and models are available at https://github.com/wpy1999/SAT.
****

Harnessing the Spatial-Temporal Attention of Diffusion Models for High-Fidelity Text-to-Image Synthesis

Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, Shiyu Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7766-7776

Diffusion-based models have achieved state-of-the-art performance on text-to-image synthesis tasks. However, one critical limitation of these models is the low fidelity of generated images with respect to the text description, such as missing objects, mismatched attributes, and mislocated objects. One key reason for such inconsistencies is the inaccurate cross-attention to text in both the spatial dimension, which controls at what pixel region an object should appear, and the temporal dimension, which controls how different levels of details are added through the denoising steps. In this paper, we propose a new text-to-image algorithm that adds explicit control over spatial-temporal cross-attention in diffusion models. We first utilize a layout predictor to predict the pixel regions for objects mentioned in the text. We then impose spatial attention control by combining the attention over the entire text description and that over the local description of the particular object in the corresponding pixel region of that object. The temporal attention control is further added by allowing the combination wei

ghts to change at each denoising step, and the combination weights are optimized to ensure high fidelity between the image and the text. Experiments show that our method generates images with higher fidelity compared to diffusion-model-based baselines without fine-tuning the diffusion model. Our code is publicly available.
****
GraphAlign: Enhancing Accurate Feature Alignment by Graph matching for Multi-Modal 3D Object Detection
Ziying Song, Haiyue Wei, Lin Bai, Lei Yang, Caiyan Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3358-3369
LiDAR and cameras are complementary sensors for 3D object detection in autonomous driving. However, it is challenging to explore the unnatural interaction between point clouds and images, and the critical factor is how to conduct feature alignment of heterogeneous modalities. Currently, many methods achieve feature alignment by projection calibration only, without considering the problem of coordinate conversion accuracy errors between sensors, leading to sub-optimal performance. In this paper, we present GraphAlign, a more accurate feature alignment strategy for 3D object detection by graph matching. Specifically, we fuse image features from a semantic segmentation encoder in the image branch and point cloud features from a 3D Sparse CNN in the LiDAR branch. To save computation, we construct the nearest neighbor relationship by calculating Euclidean distance within the subspaces that are divided into the point cloud features. Through the projection calibration between the image and point cloud, we project the nearest neighbors of point cloud features onto the image features. Then by matching the nearest neighbors with a single point cloud to multiple images, we search for a more appropriate feature alignment. In addition, we provide a self-attention module to enhance the weights of significant relations to fine-tune the feature alignment between heterogeneous modalities. Extensive experiments on nuScenes benchmark demonstrate the effectiveness and efficiency of our GraphAlign.
****
Weakly-Supervised Action Segmentation and Unseen Error Detection in Anomalous Instructional Videos
Reza Ghoddoosian, Isht Dwivedi, Nakul Agarwal, Behzad Dariush; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10128-10138
We present a novel method for weakly-supervised action segmentation and unseen error detection in anomalous instructional videos. In the absence of an appropriate dataset for this task, we introduce the Anomalous Toy Assembly (ATA) dataset, which comprises 1152 untrimmed videos of 32 participants assembling three different toys, recorded from four different viewpoints. The training set comprises 27 participants who assemble toys in an expected and consistent manner, while the test and validation sets comprise 5 participants who display sequential anomalies in their task. We introduce a weakly labeled segmentation algorithm that is a generalization of the constrained Viterbi algorithm and identifies potential anomalous moments based on the difference between future anticipation and current recognition results. The proposed method is not restricted by the training transcripts during testing, allowing for the inference of anomalous action sequences while maintaining real-time performance. Based on these segmentation results, we also introduce a baseline to detect pre-defined human errors, and benchmark results on the ATA dataset. Experiments were conducted on the ATA and CSV datasets, demonstrating that the proposed method outperforms the state-of-the-art in segmenting anomalous videos under both online and offline conditions.
****
NEMTO: Neural Environment Matting for Novel View and Relighting Synthesis of Transparent Objects
Dongqing Wang, Tong Zhang, Sabine Süsstrunk; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 317-327
We propose NEMTO, the first end-to-end neural rendering pipeline to model 3D transparent objects with complex geometry and unknown indices of refraction. Commonly used appearance modeling such as the Disney BSDF model cannot accurately addr

ess this challenging problem due to the complex light paths bending through refr actions and the strong dependency of surface appearance on illumination. With 2D images of the transparent object as input, our method is capable of high-qualit y novel view and relighting synthesis. We leverage implicit Signed Distance Func tions (SDF) to model the object geometry and propose a refraction-aware ray bend ing network to model the effects of light refraction within the object. Our ray bending network is more tolerant to geometric inaccuracies than traditional phys ically-based methods for rendering transparent objects. We provide extensive eva luations on both synthetic and real-world datasets to demonstrate our high-quali ty synthesis and the applicability of our method.
****
Geometric Viewpoint Learning with Hyper-Rays and Harmonics Encoding
Zhixiang Min, Juan Carlos Dibene, Enrique Dunn; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2023, pp. 22520-22530
Viewpoint is a fundamental modality that carries the interaction between observe rs and their environment. This paper proposes the first deep-learning framework for the viewpoint modality. The challenge in formulating learning frameworks for viewpoints resides in a suitable multimodal representation that links across th e camera viewing space and 3D environment. Traditional approaches reduce the pro blem to image analysis instances, making them computationally expensive and not adequately modelling the intrinsic geometry and environmental context of 6DoF vi ewpoints. We improve these issues in two ways. 1) We propose a generalized viewp oint representation forgoing the analysis of photometric pixels in favor of enco ded viewing ray embeddings attained from point cloud learning frameworks. 2) We propose a novel SE(3)-bijective 6D viewing ray, hyper-ray, that addresses the Do F deficiency problem of using 5DoF viewing rays representing 6DoF viewpoints. We demonstrate our approach has both efficiency and accuracy superiority over exis ting methods in novel real-world environments.
****
C2F2NeUS: Cascade Cost Frustum Fusion for High Fidelity and Generalizable Neural Surface Reconstruction
Luoyuan Xu, Tao Guan, Yuesong Wang, Wenkai Liu, Zhaojie Zeng, Junle Wang, Wei Ya ng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2023, pp. 18291-18301
There is an emerging effort to combine the two popular 3D frameworks using Multi -View Stereo (MVS) and Neural Implicit Surfaces (NIS) with a specific focus on t he few-shot / sparse view setting. In this paper, we introduce a novel integrati on scheme that combines the multi-view stereo with neural signed distance functi on representations, which potentially overcomes the limitations of both methods. MVS uses per-view depth estimation and cross-view fusion to generate accurate s urfaces, while NIS relies on a common coordinate volume. Based on this strategy, we propose to construct per-view cost frustum for finer geometry estimation, an d then fuse cross-view frustums and estimate the implicit signed distance functi ons to tackle artifacts that are due to noise and holes in the produced surface reconstruction. We further apply a cascade frustum fusion strategy to effectivel y captures global-local information and structural consistency. Finally, we appl y cascade sampling and a pseudo-geometric loss to foster stronger integration be tween the two architectures. Extensive experiments demonstrate that our method r econstructs robust surfaces and outperforms existing state-of-the-art methods.
****
Mesh2Tex: Generating Mesh Textures from Image Queries
Alexey Bokhovkin, Shubham Tulsiani, Angela Dai; Proceedings of the IEEE/CVF Inte rnational Conference on Computer Vision (ICCV), 2023, pp. 8918-8928
Remarkable advances have been achieved recently in learning neural representatio ns that characterize object geometry, while generating textured objects suitable for downstream applications and 3D rendering remains at an early stage. In part icular, reconstructing textured geometry from images of real objects is a signif icant challenge - reconstructed geometry is often inexact, making realistic text uring a significant challenge. We present Mesh2Tex, which learns a realistic obj ect texture manifold from uncorrelated collections of 3D object geometry and pho

torealistic RGB images, by leveraging a hybrid mesh-neural-field texture represe ntation. Our texture representation enables compact encoding of high-resolution textures as a neural field in the barycentric coordinate system of the mesh face s. The learned texture manifold enables effective navigation to generate an obje ct texture for a given 3D object geometry that matches to an input RGB image, wh ich maintains robustness even under challenging real-world scenarios where the m esh geometry approximates an inexact match to the underlying geometry in the RGB image. Mesh2Tex can effectively generate realistic object textures for an objec t mesh to match real images observations towards digitization of real environmen ts, significantly improving over previous state of the art.
****
USAGE: A Unified Seed Area Generation Paradigm for Weakly Supervised Semantic Se gmentation
Zelin Peng, Guanchun Wang, Lingxi Xie, Dongsheng Jiang, Wei Shen, Qi Tian; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 624-634
Seed area generation is usually the starting point of weakly supervised semantic segmentation (WSSS). Computing the Class Activation Map (CAM) from a multi-labe l classification network is the de facto paradigm for seed area generation, but CAMs generated from Convolutional Neural Networks (CNNs) and Transformers are pr one to be under- and over-activated, respectively, which makes the strategies to refine CAMs for CNNs usually inappropriate for Transformers, and vice versa. In this paper, we propose a Unified optimization paradigm for Seed Area GEneration (USAGE) for both types of networks, in which the objective function to be optim ized consists of two terms: One is a generation loss, which controls the shape o f seed areas by a temperature parameter following a deterministic principle for different types of networks; The other is a regularization loss, which ensures t he consistency between the seed areas that are generated by self-adaptive networ k adjustment from different views, to overturn false activation in seed areas. E xperimental results show that USAGE consistently improves seed area generation f or both CNNs and Transformers by large margins, e.g., outperforming state-of-the -art methods by an mIoU of 4.1% on PASCAL VOC. Moreover, based on the USAGE gene rated seed areas on Transformers, we achieve state-of-the-art WSSS results on bo th PASCAL VOC and MS COCO.
****
NeuS2: Fast Learning of Neural Implicit Surfaces for Multi-view Reconstruction
Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, Lin gjie Liu; Proceedings of the IEEE/CVF International Conference on Computer Visio n (ICCV), 2023, pp. 3295-3306
Recent methods for neural surface representation and rendering, for example NeuS , have demonstrated the remarkably high-quality reconstruction of static scenes. However, the training of NeuS takes an extremely long time (8 hours), which mak es it almost impossible to apply them to dynamic scenes with thousands of frames . We propose a fast neural surface reconstruction approach, called NeuS2, which achieves two orders of magnitude improvement in terms of acceleration without co mpromising reconstruction quality. To accelerate the training process, we parame terize a neural surface representation by multi-resolution hash encodings and pr esent a novel lightweight calculation of second-order derivatives tailored to ou r networks to leverage CUDA parallelism, achieving a factor two speed up. To fur ther stabilize and expedite training, a progressive learning strategy is propose d to optimize multi-resolution hash encodings from coarse to fine. We extend our method for fast training of dynamic scenes, with a proposed incremental trainin g strategy and a novel global transformation prediction component, which allow o ur method to handle challenging long sequences with large movements and deformat ions. Our experiments on various datasets demonstrate that NeuS2 significantly o utperforms the state-of-the-arts in both surface reconstruction accuracy and tra ining speed for both static and dynamic scenes. The code is available at our web site: https://vcai.mpi-inf.mpg.de/projects/NeuS2/.
****
Deep Feature Deblurring Diffusion for Detecting Out-of-Distribution Objects

Aming Wu, Da Chen, Cheng Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13381-13391
To promote the safe application of detectors, a task of unsupervised out-of-distribution object detection (OOD-OD) is recently proposed, whose goal is to detect unseen OOD objects without accessing any auxiliary OOD data. For this task, the challenge mainly lies in how to only leverage the known in-distribution (ID) data to detect OOD objects accurately without affecting the detection of ID objects, which can be framed as the diffusion problem for deep feature synthesis. Accordingly, such challenge could be addressed by the forward and reverse processes in the diffusion model. In this paper, we propose a new approach of Deep Feature Deblurring Diffusion (DFDD), consisting of forward blurring and reverse deblurring processes. Specifically, the forward process gradually performs Gaussian Blur on the extracted features, which is instrumental in retaining sufficient input-relevant information. By this way, the forward process could synthesize virtual OOD features that are close to the classification boundary between ID and OOD objects, which improves the performance of detecting OOD objects. During the reverse process, based on the blurred features, a dedicated deblurring model is designed to continually recover the lost details in the forward process. Both the deblurred features and original features are taken as the input for training, strengthening the discrimination ability. In the experiments, our method is evaluated on OOD-OD, open-set object detection, and incremental object detection. The significant performance gains over baselines demonstrate the superiorities of our method. The source code will be made available at: https://github.com/AmingWu/DFDD-OOD.
****
Fast Full-frame Video Stabilization with Iterative Optimization
Weiyue Zhao, Xin Li, Zhan Peng, Xianrui Luo, Xinyi Ye, Hao Lu, Zhiguo Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23534-23544
Video stabilization refers to the problem of transforming a shaky video into a visually pleasing one. The question of how to strike a good trade-off between visual quality and computational speed has remained one of the open challenges in video stabilization. Inspired by the analogy between wobbly frames and jigsaw puzzles, we propose an iterative optimization-based learning approach using synthetic datasets for video stabilization, which consists of two interacting submodules: motion trajectory smoothing and full-frame outpainting. First, we develop a two-level (coarse-to-fine) stabilizing algorithm based on the probabilistic flow field. The confidence map associated with the estimated optical flow is exploited to guide the search for shared regions through backpropagation. Second, we take a divide-and-conquer approach and propose a novel multiframe fusion strategy to render full-frame stabilized views. An important new insight brought about by our iterative optimization approach is that the target video can be interpreted as the fixed point of nonlinear mapping for video stabilization. We formulate video stabilization as a problem of minimizing the amount of jerkiness in motion trajectories, which guarantees convergence with the help of fixed-point theory. Extensive experimental results are reported to demonstrate the superiority of the proposed approach in terms of computational speed and visual quality. The code will be available on GitHub.
****
Gender Artifacts in Visual Datasets
Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4837-4848
Gender biases are known to exist within large-scale visual datasets and can be reflected or even amplified in downstream models. Many prior works have proposed methods for mitigating gender biases, often by attempting to remove gender expression information from images. To understand the feasibility and practicality of these approaches, we investigate what "gender artifacts" exist in large-scale visual datasets. We define a "gender artifact" as a visual cue correlated with gender , focusing specifically on cues that are learnable by a modern image classi

fier and have an interpretable human corollary. Through our analyses, we find th
at gender artifacts are ubiquitous in the COCO and OpenImages datasets, occurrin
g everywhere from low-level information (e.g., the mean value of the color chann
els) to higher-level image composition (e.g., pose and location of people). Furt
her, bias mitigation methods that attempt to remove gender actually remove more
information from the scene than the person. Given the prevalence of gender artif
acts, we claim that attempts to remove these artifacts from such datasets are la
rgely infeasible as certain removed artifacts may be necessary for the downstrea
m task of object recognition. Instead, the responsibility lies with researchers
and practitioners to be aware that the distribution of images within datasets is
 highly gendered and hence develop fairness-aware methods which are robust to th
ese distributional shifts across groups.
****
Learning Semi-supervised Gaussian Mixture Models for Generalized Category Discov
ery
Bingchen Zhao, Xin Wen, Kai Han; Proceedings of the IEEE/CVF International Confe
rence on Computer Vision (ICCV), 2023, pp. 16623-16633
In this paper, we address the problem of generalized category discovery (GCD), i
.e., given a set of images where part of them are labelled and the rest are not,
 the task is to automatically cluster the images in the unlabelled data, leverag
ing the information from the labelled data, while the unlabelled data contain im
ages from the labelled classes and also new ones. GCD is similar to semi-supervi
sed learning (SSL) but is more realistic and challenging, as SSL assumes all the
 unlabelled images are from the same classes as the labelled ones.
 We also do not assume the class number in the unlabelled data is known a-priori
, making the GCD problem even harder.
 To tackle the problem of GCD without knowing the class number, we propose an EM
-like framework that alternates between representation learning and class number
 estimation. We propose a semi-supervised variant of the Gaussian Mixture Model
(GMM) with a stochastic splitting and merging mechanism to dynamically determine
 the prototypes by examining the cluster compactness and separability. With thes
e prototypes, we leverage prototypical contrastive learning for representation l
earning on the partially labelled data subject to the constraints imposed by the
 labelled data. Our framework alternates between these two steps until convergen
ce. The cluster assignment for an unlabelled instance can then be retrieved by i
dentifying its nearest prototype. We comprehensively evaluate our framework on b
oth generic image classification datasets and challenging fine-grained object re
cognition datasets, achieving state-of-the-art performance.
****
SuS-X: Training-Free Name-Only Transfer of Vision-Language Models
Vishaal Udandarao, Ankush Gupta, Samuel Albanie; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 2725-2736
Contrastive Language-Image Pre-training (CLIP) has emerged as a simple yet effec
tive way to train large-scale vision-language models. CLIP demonstrates impressi
ve zero-shot classification and retrieval performance on diverse downstream task
s. However, to leverage its full potential, fine-tuning still appears to be nece
ssary. Fine-tuning the entire CLIP model can be resource-intensive and unstable.
 Moreover, recent methods that aim to circumvent this need for fine-tuning still
 require access to images from the target distribution. In this paper, we pursue
 a different approach and explore the regime of training-free "name-only transfe
r" in which the only knowledge we possess about downstream tasks comprises the n
ames of downstream target categories. We propose a novel method, SuS-X, consisti
ng of two key building blocks--"SuS" and "TIP-X", that requires neither intensiv
e fine-tuning nor costly labelled data. SuS-X achieves state-of-the-art (SoTA) z
ero-shot classification results on 19 benchmark datasets. We further show the ut
ility of TIP-X in the training-free few-shot setting, where we again achieve SoT
A results over strong training-free baselines.
****
Rethinking Point Cloud Registration as Masking and Reconstruction
Guangyan Chen, Meiling Wang, Li Yuan, Yi Yang, Yufeng Yue; Proceedings of the IE

EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17717-17727
Point cloud registration is essential in computer vision and robotics. In this paper, a critical observation is made that the invisible parts of each point cloud can be directly utilized as inherent masks, and the aligned point cloud pair can be regarded as the reconstruction target. Motivated by this observation, we rethink the point cloud registration problem as a masking and reconstruction task. To this end, a generic and concise auxiliary training network, the Masked Reconstruction Auxiliary Network (MRA), is proposed. The MRA reconstructs the complete point cloud by separately using the encoded features of each point cloud obtained from the backbone, guiding the contextual features in the backbone to capture fine-grained geometric details and the overall structures of point cloud pairs. Unlike recently developed high-performing methods that incorporate specific encoding methods into transformer models, which sacrifice versatility and introduce significant computational complexity during the inference process, our MRA can be easily inserted into other methods to further improve registration accuracy. Additionally, the MRA is detached after training, thereby avoiding extra computational complexity during the inference process. Building upon the MRA, we present a novel transformer-based method, the Masked Reconstruction Transformer (MRT), which achieves both precise and efficient alignment using standard transformers. Extensive experiments conducted on the 3DMatch, ModelNet40, and KITTI datasets demonstrate the superior performance of our MRT over state-of-the-art methods, and the efficiency of the MRA in improving registration accuracy.
****
Beating Backdoor Attack at Its Own Game
Min Liu, Alberto Sangiovanni-Vincentelli, Xiangyu Yue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4620-4629
Deep neural networks (DNNs) are vulnerable to backdoor attack, which does not affect the network's performance on clean data but would manipulate the network behavior once a trigger pattern is added.
Existing defense methods have greatly reduced attack success rate, but their prediction accuracy on clean data still lags behind a clean model by a large margin.
Inspired by the stealthiness and effectiveness of backdoor attack, we propose a simple but highly effective defense framework which injects non-adversarial backdoors targeting poisoned samples.
Following the general steps in backdoor attack, we detect a small set of suspected samples and then apply a poisoning strategy to them.
The non-adversarial backdoor, once triggered, suppresses the attacker's backdoor on poisoned data, but has limited influence on clean data.
The defense can be carried out during data preprocessing, without any modification to the standard end-to-end training pipeline.
We conduct extensive experiments on multiple benchmarks with different architectures and representative attacks.
Results demonstrate that our method achieves state-of-the-art defense effectiveness with by far the lowest performance drop on clean data.
Considering the surprising defense ability displayed by our framework, we call for more attention to utilizing backdoor for backdoor defense.
Code is available at https://github.com/damianliumin/non-adversarial_backdoor.
****
Introducing Language Guidance in Prompt-based Continual Learning
Muhammad Gul Zain Ali Khan, Muhammad Ferjad Naeem, Luc Van Gool, Didier Stricker, Federico Tombari, Muhammad Zeshan Afzal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11463-11473
Continual Learning aims to learn a single model on a sequence of tasks without having access to data from previous tasks. The biggest challenge in the domain still remains catastrophic forgetting: a loss in performance on seen classes of earlier tasks. Some existing methods rely on an expensive replay buffer to store a chunk of data from previous tasks. This, while promising, becomes expensive when the number of tasks becomes large or data can not be stored for privacy reasons. As an alternative, prompt-based methods have been proposed that store the tas

k information in a learnable prompt pool. This prompt pool instructs a frozen im
age encoder on how to solve each task. While the model faces a disjoint set of c
lasses in each task in this setting, we argue that these classes can be encoded
to the same embedding space of a pre-trained language encoder. In this work, we
propose Language Guidance for Prompt-based Continual Learning (LGCL) as a plug-i
n for prompt-based methods. LGCL is model agnostic and introduces language guida
nce at the task level in the prompt pool and at the class level on the output fe
ature of the vision encoder. We show with extensive experimentation that LGCL co
nsistently improves the performance of prompt-based continual learning methods t
o set a new state-of-the-art. LGCL achieves these performance improvements witho
ut needing any additional learnable parameters.
****
Invariant Training 2D-3D Joint Hard Samples for Few-Shot Point Cloud Recognition
Xuanyu Yi, Jiajun Deng, Qianru Sun, Xian-Sheng Hua, Joo-Hwee Lim, Hanwang Zhang;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2023, pp. 14463-14474
We tackle the data scarcity challenge in few-shot point cloud recognition of 3D
objects by using a joint prediction from a conventional 3D model and a well-pret
rained 2D model. Surprisingly, such an ensemble, though seems trivial, has hardl
y been shown effective in recent 2D-3D models. We find out the crux is the less
effective training for the "joint hard samples", which have high confidence pred
iction on different wrong labels, implying that the 2D and 3D models do not coll
aborate well. To this end, our proposed invariant training strategy, called INVJ
OINT, does not only emphasize the training more on the hard samples, but also se
eks the invariance between the conflicting 2D and 3D ambiguous predictions. INVJ
OINT can learn more collaborative 2D and 3D representations for better ensemble.
 Extensive experiments on 3D shape classification with widely-adopted ModelNet10
/40, ScanObjectNN and Toys4K, and shape retrieval with ShapeNet-Core validate th
e superiority of our INVJOINT.
****
EigenPlaces: Training Viewpoint Robust Models for Visual Place Recognition
Gabriele Berton, Gabriele Trivigno, Barbara Caputo, Carlo Masone; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1108
0-11090
Visual Place Recognition is a task that aims to predict the place of an image (c
alled query) based solely on its visual features.
 This is typically done through image retrieval, where the query is matched to t
he most similar images from a large database of geotagged photos, using learned
global descriptors.
 A major challenge in this task is recognizing places seen from different viewpo
ints. To overcome this limitation, we propose a new method, called EigenPlaces,
to train our neural network on images from different point of views, which embed
s viewpoint robustness into the learned global descriptors. The underlying idea
is to cluster the training data so as to explicitly present the model with diffe
rent views of the same points of interest. The selection of this points of inter
est is done without the need for extra supervision.
 We then present experiments on the most comprehensive set of datasets in litera
ture, finding that EigenPlaces is able to outperform previous state of the art o
n the majority of datasets, while requiring 60% less GPU memory for training and
 using 50% smaller descriptors.
 The code and trained models for EigenPlaces are available at https://github.com
/gmberton/EigenPlaces, while results with any other baseline can be computed wit
h the codebase at https://github.com/gmberton/auto_VPR.
****
Do DALL-E and Flamingo Understand Each Other?
Hang Li, Jindong Gu, Rajat Koner, Sahand Sharifzadeh, Volker Tresp; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19
99-2010
The field of multimodal research focusing on the comprehension and creation of b
oth images and text has witnessed significant strides. This progress is exemplif

ied by the emergence of sophisticated models dedicated to image captioning at scale, such as the notable Flamingo model and text-to-image generative models, with DALL-E serving as a prominent example. An interesting question worth exploring in this domain is whether Flamingo and DALL-E understand each other. To study this question, we propose a reconstruction task where Flamingo generates a description for a given image and DALL-E uses this description as input to synthesize a new image. We argue that these models understand each other if the generated image is similar to the given image. Specifically, we study the relationship between the quality of the image reconstruction and that of the text generation. We find that an optimal description of an image is one that gives rise to a generated image similar to the original one. The finding motivates us to propose a unified framework to finetune the text-to-image and image-to-text models. Concretely, the reconstruction part forms a regularization loss to guide the tuning of the models. Extensive experiments on multiple datasets with different image captioning and image generation models validate our findings and demonstrate the effectiveness of our proposed unified framework. As DALL-E and Flamingo are not publicly available, we use Stable Diffusion and BLIP in the remaining work. Project website: https://dalleflamingo.github.io.

****

CIRI: Curricular Inactivation for Residue-aware One-shot Video Inpainting

Weiying Zheng, Cheng Xu, Xuemiao Xu, Wenxi Liu, Shengfeng He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13012-13022

Video inpainting aims at filling in missing regions of a video. However, when dealing with dynamic scenes with camera or object movements, annotating the inpainting target becomes laborious and impractical. In this paper, we resolve the one-shot video inpainting problem in which only one annotated first frame is provided. A naive solution is to propagate the initial target to the other frames with techniques like object tracking. In this context, the main obstacles are the unreliable propagation and the partially inpainted artifacts due to the inaccurate mask. For the former problem, we propose curricular inactivation to replace the hard masking mechanism for indicating the inpainting target, which is robust to erroneous predictions in long-term video inpainting. For the latter, we explore the properties of inpainting residue and present an online residue removal method in an iterative detect-and-refine manner. Extensive experiments on several real-world datasets demonstrate the quantitative and qualitative superiorities of our proposed method in one-shot video inpainting. More importantly, our method is extremely flexible that can be integrated with arbitrary traditional inpainting models, activating them to perform the reliable one-shot video inpainting task. Video demonstrations can be found in our supplement, and our code can be found at https://github.com/Arise-zwy/CIRI.

****

Protoype-based Dataset Comparison

Nanne van Noord; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1944-1954

Dataset summarisation is a fruitful approach to dataset inspection. However, when applied to a single dataset the discovery of visual concepts is restricted to those most prominent. We argue that a comparative approach can expand upon this paradigm to enable richer forms of dataset inspection that go beyond the most prominent concepts. To enable dataset comparison we present a module that learns concept-level prototypes across datasets. We leverage self-supervised learning to discover these prototypes without supervision, and we demonstrate the benefits of our approach in two case-studies. Our findings show that dataset comparison extends dataset inspection and we hope to encourage more works in this direction. Code and usage instructions available at https://github.com/Nanne/ProtoSim

****

FreeCOS: Self-Supervised Learning from Fractals and Unlabeled Images for Curvilinear Object Segmentation

Tianyi Shi, Xiaohuan Ding, Liang Zhang, Xin Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 876-886

Curvilinear object segmentation is critical for many applications. However, manu
ally annotating curvilinear objects is very time-consuming and error-prone, yiel
ding insufficiently available annotated datasets for existing supervised methods
 and domain adaptation methods. This paper proposes a self-supervised curvilinea
r object segmentation method (FreeCOS) that learns robust and distinctive featur
es from fractals and unlabeled images. The key contributions include a novel Fra
ctal-FDA synthesis (FFS) module and a geometric information alignment (GIA) appr
oach. FFS generates curvilinear structures based on the parametric Fractal L-sys
tem and integrates the generated structures into unlabeled images to obtain synt
hetic training images via Fourier Domain Adaptation. GIA reduces the intensity d
ifferences between the synthetic and unlabeled images by comparing the intensity
 order of a given pixel to the values of its nearby neighbors. Such image alignm
ent can explicitly remove the dependency on absolute intensity values and enhanc
e the inherent geometric characteristics which are common in both synthetic and
real images. In addition, GIA aligns features of synthetic and real images via t
he prediction space adaptation loss (PSAL) and the curvilinear mask contrastive
loss (CMCL). Extensive experimental results on four public datasets, i.e., XCAD,
 DRIVE, STARE and CrackTree demonstrate that our method outperforms the state-of
-the-art unsupervised methods, self-supervised methods and traditional methods b
y a large margin. The source code of this work is available at https://github.co
m/TY-Shi/FreeCOS.
****
Generating Dynamic Kernels via Transformers for Lane Detection
Ziye Chen, Yu Liu, Mingming Gong, Bo Du, Guoqi Qian, Kate Smith-Miles; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
 6835-6844
State-of-the-art lane detection methods often rely on specific knowledge about l
anes -- such as straight lines and parametric curves -- to detect lane lines. Wh
ile the specific knowledge can ease the modeling process, it poses challenges in
 handling lane lines with complex topologies (e.g., dense, forked, curved, etc.)
. Recently, dynamic convolution-based methods have shown promising performance b
y utilizing the features from some key locations of a lane line, such as the sta
rting point, as convolutional kernels, and convoluting them with the whole featu
re map to detect lane lines. While such methods reduce the reliance on specific
knowledge, the kernels computed from the key locations fail to capture the lane
line's global structure due to its long and thin structure, leading to inaccurat
e detection of lane lines with complex topologies. In addition, the kernels resu
lting from the key locations are sensitive to occlusion and lane intersections.
To overcome these limitations, we propose a transformer-based dynamic kernel gen
eration architecture for lane detection. It utilizes a transformer to generate d
ynamic convolutional kernels for each lane line in the input image, and then det
ect these lane lines with dynamic convolution. Compared to the kernels generated
 from the key locations of a lane line, the kernels generated with the transform
er can capture the lane line's global structure from the whole feature map, enab
ling them to effectively handle occlusions and lane lines with complex topologie
s. We evaluate our method on three lane detection benchmarks, and the results de
monstrate its state-of-the-art performance. Specifically, our method achieves an
 F1 score of 63.40 on OpenLane and 88.47 on CurveLanes, surpassing the state of
the art by 4.30 and 2.37 points, respectively.
****
RSFNet: A White-Box Image Retouching Approach using Region-Specific Color Filter
s
Wenqi Ouyang, Yi Dong, Xiaoyang Kang, Peiran Ren, Xin Xu, Xuansong Xie; Proceedi
ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp
. 12160-12169
Retouching images is an essential aspect of enhancing the visual appeal of photo
s. Although users often share common aesthetic preferences, their retouching met
hods may vary based on their individual preferences. Therefore, there is a need
for white-box approaches that produce satisfying results and enable users to con
veniently edit their images simultaneously. Recent white-box retouching methods

rely on cascaded global filters that provide image-level filter arguments but cannot perform fine-grained retouching. In contrast, colorists typically employ a divide-and-conquer approach, performing a series of region-specific fine-grained enhancements when using traditional tools like Davinci Resolve. We draw on this insight to develop a white-box framework for photo retouching using parallel region-specific filters, called RSFNet. Our model generates filter arguments (e.g., saturation, contrast, hue) and attention maps of regions for each filter simultaneously. Instead of cascading filters, RSFNet employs linear summations of filters, allowing for a more diverse range of filter classes that can be trained more easily. Our experiments demonstrate that RSFNet achieves state-of-the-art results, offering satisfying aesthetic appeal and increased user convenience for editable white-box retouching.
****
Tem-Adapter: Adapting Image-Text Pretraining for Video Question Answer
Guangyi Chen, Xiao Liu, Guangrun Wang, Kun Zhang, Philip H.S. Torr, Xiao-Ping Zhang, Yansong Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13945-13955
Video-language pre-trained models have shown remarkable success in guiding video question-answering (VideoQA) tasks. However, due to the length of video sequences, training large-scale video-based models incurs considerably higher costs than training image-based ones. This motivates us to leverage the knowledge from image-based pretraining, despite the obvious gaps between image and video domains. To bridge these gaps, in this paper, we propose Tem-Adapter, which enables the learning of temporal dynamics and complex semantics by a visual Temporal Aligner and a textual Semantic Aligner. Unlike conventional pretrained knowledge adaptation methods that only concentrate on the downstream task objective, the Temporal Aligner introduces an extra language-guided autoregressive task aimed at facilitating the learning of temporal dependencies, with the objective of predicting future states based on historical clues and language guidance that describes event progression. Besides, to reduce the semantic gap and adapt the textual representation for better event description, we introduce a Semantic Aligner that first designs a template to fuse question and answer pairs as event descriptions and then learns a Transformer decoder with the whole video sequence as guidance for refinement. We evaluate Tem-Adapter and different pre-train transferring methods on two VideoQA benchmarks, and the significant performance improvement demonstrates the effectiveness of our method.
****
Boosting Long-tailed Object Detection via Step-wise Learning on Smooth-tail Data
Na Dong, Yongqiang Zhang, Mingli Ding, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6940-6949
Real-world data tends to follow a long-tailed distribution, where the class imbalance results in dominance of the head classes during training. In this paper, we propose a frustratingly simple but effective step-wise learning framework to gradually enhance the capability of the model in detecting all categories of long-tailed datasets. Specifically, we build smooth-tail data where the long-tailed distribution of categories decays smoothly to correct the bias towards head classes. We pre-train a model on the whole long-tailed data to preserve discriminability between all categories. We then fine-tune the class-agnostic modules of the pre-trained model on the head class dominant replay data to get a head class expert model with improved decision boundaries from all categories. Finally, we train a unified model on the tail class dominant replay data while transferring knowledge from the head class expert model to ensure accurate detection of all categories. Extensive experiments on long-tailed datasets LVIS v0.5 and LVIS v1.0 demonstrate the superior performance of our method, where we can improve the AP with ResNet-50 backbone from 27.0% to 30.3% AP, and especially for the rare categories from 15.5% to 24.9% AP. Our best model using ResNet-101 backbone can achieve 30.7% AP, which suppresses all existing detectors using the same backbone.
****
Talking Head Generation with Probabilistic Audio-to-Visual Diffusion Priors
Zhentao Yu, Zixin Yin, Deyu Zhou, Duomin Wang, Finn Wong, Baoyuan Wang; Proceedi

ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7645-7655

We introduce a novel framework for one-shot audio-driven talking head generation. Unlike prior works that require additional driving sources for controlled synthesis in a deterministic manner, we instead sample all holistic lip-irrelevant facial motions (i.e. pose, expression, blink, gaze, etc.) to semantically match the input audio while still maintaining both the photo-realism of audio-lip synchronization and overall naturalness. This is achieved by our newly proposed audio-to-visual diffusion prior trained on top of the mapping between audio and non-lip representations. Thanks to the probabilistic nature of the diffusion prior, one big advantage of our framework is it can synthesize diverse facial motion sequences given the same audio clip, which is quite user-friendly for many real applications. Through comprehensive evaluations of public benchmarks, we conclude that (1) our diffusion prior outperforms auto-regressive prior significantly on all the concerned metrics; (2) our overall system is competitive with prior works in terms of audio-lip synchronization but can effectively sample rich and natural-looking lip-irrelevant facial motions while still semantically harmonized with the audio input.

****

Learning Cross-Modal Affinity for Referring Video Object Segmentation Targeting Limited Samples

Guanghui Li, Mingqi Gao, Heng Liu, Xiantong Zhen, Feng Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2684-2693

Referring video object segmentation (RVOS), as a supervised learning task, relies on sufficient annotated data for a given scene. However, in more realistic scenarios, only minimal annotations are available for a new scene, which poses significant challenges to existing RVOS methods. With this in mind, we propose a simple yet effective model with a newly designed cross-modal affinity (CMA) module based on a Transformer architecture. The CMA module builds multimodal affinity with a few samples, thus quickly learning new semantic information, and enabling the model to adapt to different scenarios. Since the proposed method targets limited samples for new scenes, we generalize the problem as - few-shot referring video object segmentation (FS-RVOS). To foster research in this direction, we build up a new FS-RVOS benchmark based on currently available datasets. The benchmark covers a wide range and includes multiple situations, which can maximally simulate real-world scenarios. Extensive experiments show that our model adapts well to different scenarios with only a few samples, reaching state-of-the-art performance on the benchmark. On Mini-Ref-YouTube-VOS, our model achieves an average performance of 53.1 J and 54.8 F, which are 10% better than the baselines. Furthermore, we show impressive results of 77.7 J and 74.8 F on Mini-Ref-SAIL-VOS, which are significantly better than the baselines. Code is publicly available at https://github.com/hengliusky/Few_shot_RVOS.

****

Human Part-wise 3D Motion Context Learning for Sign Language Recognition

Taeryung Lee, Yeonguk Oh, Kyoung Mu Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20740-20750

In this paper, we propose P3D, the human part-wise motion context learning framework for sign language recognition. Our main contributions lie in two dimensions: learning the part-wise motion context and employing the pose ensemble to utilize 2D and 3D pose jointly. First, our empirical observation implies that part-wise context encoding benefits the performance of sign language recognition. While previous methods of sign language recognition learned motion context from the sequence of the entire pose, we argue that such methods cannot exploit part-specific motion context. In order to utilize part-wise motion context, we propose the alternating combination of a part-wise encoding Transformer (PET) and a whole-body encoding Transformer (WET). PET encodes the motion contexts from a part sequence, while WET merges them into a unified context. By learning part-wise motion context, our P3D achieves superior performance on WLASL compared to previous state-of-the-art methods. Second, our framework is the first to ensemble 2D and 3D

poses for sign language recognition. Since the 3D pose holds rich motion contex t and depth information to distinguish the words, our P3D outperformed the previ ous state-of-the-art methods employing a pose ensemble.
****
Remembering Normality: Memory-guided Knowledge Distillation for Unsupervised Ano maly Detection
Zhihao Gu, Liang Liu, Xu Chen, Ran Yi, Jiangning Zhang, Yabiao Wang, Chengjie Wa ng, Annan Shu, Guannan Jiang, Lizhuang Ma; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2023, pp. 16401-16409
Knowledge distillation (KD) has been widely explored in unsupervised anomaly det ection (AD). The student is assumed to constantly produce representations of typ ical patterns within trained data, named "normality", and the representation dis crepancy between the teacher and student model is identified as anomalies. Howev er, it suffers from the "normality forgetting" issue. Trained on anomaly-free da ta, the student still well reconstructs anomalous representations for anomalies and is sensitive to fine patterns in normal data, which also appear in training. To mitigate this issue, we introduce a novel Memory-guided Knowledge-Distillati on (MemKD) framework that adaptively modulates the normality of student features in detecting anomalies. Specifically, we first propose a normality recall memor y (NR Memory) to strengthen the normality of student-generated features by recal ling the stored normal information. In this sense, representations will not pres ent anomalies and fine patterns will be well described. Subsequently, we employ a normality embedding learning strategy to promote information learning for the NR Memory. It constructs a normal exemplar set so that the NR Memory can memoriz e prior knowledge in anomaly-free data and later recall them from the query feat ure. Consequently, comprehensive experiments demonstrate that the proposed MemKD achieves promising results on five benchmarks, i.e., MVTec AD, VisA, MPDD, MVTe c 3D-AD, and Eyecandies.
****
Coordinate Quantized Neural Implicit Representations for Multi-view Reconstructi on
Sijia Jiang, Jing Hua, Zhizhong Han; Proceedings of the IEEE/CVF International C onference on Computer Vision (ICCV), 2023, pp. 18358-18369
In recent years, huge progress has been made on learn- ing neural implicit repre sentations from multi-view images for 3D reconstruction. As an additional input complement- ing coordinates, using sinusoidal functions as positional encodings plays a key role in revealing high frequency de- tails with coordinate-based neu ral networks. However, high frequency positional encodings make the optimization un- stable, which results in noisy reconstructions and artifacts in empty space . To resolve this issue in a general sense, we introduce to learn neural implici t representations with quantized coordinates, which reduces the uncertainty and ambiguity in the field during optimization. Instead of con- tinuous coordinates, we discretize continuous coordinates into discrete coordinates using nearest in terpolation among quantized coordinates which are obtained by discretizing the f ield in an extremely high resolution. We use discrete coordinates and their posi tional encodings to learn implicit functions through volume rendering. This sign ificantly re- duces the variations in the sample space, and triggers more multi- view consistency constraints on intersections of rays from different views, whic h enables to infer implicit function in a more effective way. Our quantized coor dinates do not bring any computational burden, and can seamlessly work upon the latest methods. Our evaluations under the widely used benchmarks show our superi ority over the state-of-the- art. Our code is available at https://github.com/ M achinePerceptionLab/CQ-NIR.
****
Unleashing the Potential of Spiking Neural Networks with Dynamic Confidence
Chen Li, Edward G Jones, Steve Furber; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13350-13360
This paper presents a new methodology to alleviate the fundamental trade-off bet ween accuracy and latency in spiking neural networks (SNNs). The approach involv es decoding confidence information over time from the SNN outputs and using it t

o develop a decision-making agent that can dynamically determine when to terminate each inference.

The proposed method, Dynamic Confidence, provides several significant benefits to SNNs. 1. It can effectively optimize latency dynamically at runtime, setting it apart from many existing low-latency SNN algorithms. Our experiments on CIFAR-10 and ImageNet datasets have demonstrated an average 40% speedup across eight different settings after applying Dynamic Confidence. 2. The decision-making agent in Dynamic Confidence is straightforward to construct and highly robust in parameter space, making it extremely easy to implement. 3. The proposed method enables visualizing the potential of any given SNN, which sets a target for current SNNs to approach. For instance, if an SNN can terminate at the most appropriate time point for each input sample, a ResNet-50 SNN can achieve an accuracy as high as 82.47% on ImageNet within just 4.71 time steps on average. Unlocking the potential of SNNs needs a highly-reliable decision-making agent to be constructed and fed with a high-quality estimation of ground truth. In this regard, Dynamic Confidence represents a meaningful step toward realizing the potential of SNNs.
****
TeD-SPAD: Temporal Distinctiveness for Self-Supervised Privacy-Preservation for Video Anomaly Detection
Joseph Fioresi, Ishan Rajendrakumar Dave, Mubarak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13598-13609
Video anomaly detection (VAD) without human monitoring is a complex computer vision task that can have a positive impact on society if implemented successfully. While recent advances have made significant progress in solving this task, most existing approaches overlook a critical real-world concern: privacy. With the increasing popularity of artificial intelligence technologies, it becomes crucial to implement proper AI ethics into their development. Privacy leakage in VAD allows models to pick up and amplify unnecessary biases related to people's personal information, which may lead to undesirable decision making. In this paper, we propose TeD-SPAD, a privacy-aware video anomaly detection framework that destroys visual private information in a self-supervised manner. In particular, we propose the use of a temporally-distinct triplet loss to promote temporally discriminative features, which complements current weakly-supervised VAD methods. Using TeD-SPAD, we achieve a positive trade-off between privacy protection and utility anomaly detection performance on three popular weakly supervised VAD datasets: UCF-Crime, XD-Violence, and ShanghaiTech. Our proposed anonymization model reduces private attribute prediction by 32.25% while only reducing frame-level ROC AUC on the UCF-Crime anomaly detection dataset by 3.69%.
****
MAS: Towards Resource-Efficient Federated Multiple-Task Learning
Weiming Zhuang, Yonggang Wen, Lingjuan Lyu, Shuai Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23414-23424
Federated learning (FL) is an emerging distributed machine learning method that empowers in-situ model training on decentralized edge devices. However, multiple simultaneous FL tasks could overload resource-constrained devices. In this work, we propose the first FL system to effectively coordinate and train multiple simultaneous FL tasks. We first formalize the problem of training simultaneous FL tasks. Then, we present our new approach, MAS (Merge and Split), to optimize the performance of training multiple simultaneous FL tasks. MAS starts by merging FL tasks into an all-in-one FL task with a multi-task architecture. After training for a few rounds, MAS splits the all-in-one FL task into two or more FL tasks by using the affinities among tasks measured during the all-in-one training. It then continues training each split of FL tasks based on model parameters from the all-in-one training. Extensive experiments demonstrate that MAS outperforms other methods while reducing training time by 2x and reducing energy consumption by 40%. We hope this work will inspire the community to further study and optimize training simultaneous FL tasks.
****
Bridging Cross-task Protocol Inconsistency for Distillation in Dense Object Dete

ction
Longrong Yang, Xianpan Zhou, Xuewei Li, Liang Qiao, Zheyang Li, Ziwei Yang, Gaoa ng Wang, Xi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17175-17184
Knowledge distillation (KD) has shown potential for learning compact models in d ense object detection. However, the commonly used softmax-based distillation ign ores the absolute classification scores for individual categories. Thus, the opt imum of the distillation loss does not necessarily lead to the optimal student c lassification scores for dense object detectors. This cross-task protocol incons istency is critical, especially for dense object detectors, since the foreground categories are extremely imbalanced. To address the issue of protocol differenc es between distillation and classification, we propose a novel distillation meth od with cross-task consistent protocols, tailored for the dense object detection . For classification distillation, we address the cross-task protocol inconsiste ncy problem by formulating the classification logit maps in both teacher and stu dent models as multiple binary-classification maps and applying a binary-classif ication distillation loss to each map. For localization distillation, we design an IoU-based Localization Distillation Loss that is free from specific network s tructures and can be compared with existing localization distillation losses. Ou r proposed method is simple but effective, and experimental results demonstrate its superiority over existing methods.
****
Divide and Conquer: a Two-Step Method for High Quality Face De-identification wi th Model Explainability
Yunqian Wen, Bo Liu, Jingyi Cao, Rong Xie, Li Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5148-5157
Face de-identification involves concealing the true identity of a face while ret aining other facial characteristics. Current target-generic methods typically di sentangle identity features in the latent space, using adversarial training to b alance privacy and utility. However, this pattern often leads to a trade-off bet ween privacy and utility, and the latent space remains difficult to explain. To address these issues, we propose IDeudemon, which employs a "divide and conquer" strategy to protect identity and preserve utility step by step while maintainin g good explainability. In Step I, we obfuscate the 3D disentangled ID code calcu lated by a parametric NeRF model to protect identity. In Step II, we incorporate visual similarity assistance and train a GAN with adjusted losses to preserve i mage utility. Thanks to the powerful 3D prior and delicate generative designs, o ur approach could protect the identity naturally, produce high quality details a nd is robust to different poses and expressions. Extensive experiments demonstra te that the proposed IDeudemon outperforms previous state-of-the-art methods.
****
HiTeA: Hierarchical Temporal-Aware Video-Language Pre-training
Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, Fei Huang; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15405-15416
Video-language pre-training has advanced the performance of various downstream v ideo-language tasks. However, most previous methods directly inherit or adapt ty pical image-language pre-training paradigms to video-language pre-training, thus not fully exploiting the unique characteristic of video, i.e., temporal. In thi s paper, we propose a Hierarchical Temporal-Aware video-language pre-training fr amework, HiTeA, with two novel pre-training tasks for yielding temporal-aware mu lti-modal representation with cross-modal fine-grained temporal moment informati on and temporal contextual relations between video-text multi-modal pairs. First , we propose a cross-modal moment exploration task to explore moments in videos by mining the paired texts, which results in detailed video moment representatio n. Then, based on the learned detailed moment representations, the inherent temp oral contextual relations are captured by aligning video-text pairs as a whole i n different time resolutions with multi-modal temporal relation exploration task . Furthermore, we introduce the shuffling test to evaluate the temporal reliance of datasets and video-language pre-training models. We achieve state-of-the-art

results on 15 well-established video-language understanding and generation task
s, especially on temporal-oriented datasets (e.g., SSv2-Template and SSv2-Label)
with 8.6% and 11.1% improvement respectively. HiTeA also demonstrates strong ge
neralization ability when directly transferred to downstream tasks in a zero-sho
t manner.
****

VAPCNet: Viewpoint-Aware 3D Point Cloud Completion
Zhiheng Fu, Longguang Wang, Lian Xu, Zhiyong Wang, Hamid Laga, Yulan Guo, Farid
Boussaid, Mohammed Bennamoun; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2023, pp. 12108-12118
Most existing learning-based 3D point cloud completion methods ignore the fact t
hat the completion process is highly coupled with the viewpoint of a partial sca
n. However, the various viewpoints of incompletely scanned objects in real-world
applications are normally unknown and directly estimating the viewpoint of each
incomplete object is usually time-consuming and leads to huge annotation cost.
In this paper, we thus propose an unsupervised viewpoint representation learning
scheme for 3D point cloud completion without explicit viewpoint estimation. To
be specific, we learn abstract representations of partial scans to distinguish v
arious viewpoints in the representation space rather than the explicit estimatio
n in the 3D space. We also introduce a Viewpoint-Aware Point cloud Completion Ne
twork (VAPCNet) with flexible adaption to various viewpoints based on the learne
d
representations. The proposed viewpoint representation learning scheme can extr
act discriminative representations to obtain accurate viewpoint information. Rep
orted experiments on two popular public datasets show that our VAPCNet achieves
state-of-the-art performance for the point cloud completion task. Source code is
available at https://github.com/FZH92128/VAPCNet.
****

Set-level Guidance Attack: Boosting Adversarial Transferability of Vision-Langua
ge Pre-training Models
Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, Feng Zheng; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 102-111
Vision-language pre-training (VLP) models have shown vulnerability to adversaria
l examples in multimodal tasks. Furthermore, malicious adversaries can be delibe
rately transferred to attack other black-box models. However, existing work has
mainly focused on investigating white-box attacks. In this paper, we present the
first study to investigate the adversarial transferability of recent VLP models
. We observe that existing methods exhibit much lower transferability, compared
to the strong attack performance in white-box settings. The transferability degr
adation is partly caused by the under-utilization of cross-modal interactions. P
articularly, unlike unimodal learning, VLP models rely heavily on cross-modal in
teractions and the multimodal alignments are many-to-many, e.g., an image can be
described in various natural languages. To this end, we propose a highly transf
erable Set-level Guidance Attack (SGA) that thoroughly leverages modality intera
ctions and incorporates alignment-preserving augmentation with cross-modal guida
nce. Experimental results demonstrate that SGA could generate adversarial exampl
es that can strongly transfer across different VLP models on multiple downstream
vision-language tasks. On image-text retrieval, SGA significantly enhances the
attack success rate for transfer attacks from ALBEF to TCL by a large margin (at
least 9.78% and up to 30.21%), compared to the state-of-the-art. Our code is av
ailable at https://github.com/Zoky-2020/SGA.
****

AutoSynth: Learning to Generate 3D Training Data for Object Point Cloud Registra
tion
Zheng Dang, Mathieu Salzmann; Proceedings of the IEEE/CVF International Conferen
ce on Computer Vision (ICCV), 2023, pp. 9009-9019
In the current deep learning paradigm, the amount and quality of training data a
re as critical as the network architecture and its training details. However, co
llecting, processing, and annotating real data at scale is difficult, expensive,

and time-consuming, particularly for tasks such as 3D object registration. While synthetic datasets can be created, they require expertise to design and include a limited number of categories. In this paper, we introduce a new approach called AutoSynth, which automatically generates 3D training data for point cloud registration. Specifically, AutoSynth automatically curates an optimal dataset by exploring a search space encompassing millions of potential datasets with diverse 3D shapes at a low cost. To achieve this, we generate synthetic 3D datasets by assembling shape primitives, and develop a meta-learning strategy to search for the best training data for 3D registration on real point clouds. For this search to remain tractable, we replace the point cloud registration network with a much smaller surrogate network, leading to a 4056.43 times speedup. We demonstrate the generality of our approach by implementing it with two different point cloud registration networks, BPNet and IDAM. Our results on TUD-L, LINEMOD, and Occluded-LINEMOD evidence that a neural network trained on our searched dataset yields consistently better performance than the same one trained on the widely used ModelNet40 dataset.
****
Multimodal Distillation for Egocentric Action Recognition
Gorjan Radevski, Dusan Grujicic, Matthew Blaschko, Marie-Francine Moens, Tinne Tuytelaars; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5213-5224
The focal point of egocentric video understanding is modelling hand-object interactions. Standard models, e.g. CNNs or Vision Transformers, which receive RGB frames as input perform well, however, their performance improves further by employing additional input modalities that provide complementary cues, such as object detections, optical flow, audio, etc. The added complexity of the modality-specific modules, on the other hand, makes these models impractical for deployment. The goal of this work is to retain the performance of such a multimodal approach, while using only the RGB frames as input at inference time. We demonstrate that for egocentric action recognition on the Epic-Kitchens and the Something-Something datasets, students which are taught by multimodal teachers tend to be more accurate and better calibrated than architecturally equivalent models trained on ground truth labels in a unimodal or multimodal fashion. We further present a principled multimodal knowledge distillation framework, allowing us to deal with issues which occur when applying multimodal knowledge distillation in a naive manner. Lastly, we demonstrate the achieved reduction in computational complexity, and show that our approach maintains higher performance with the reduction of the number of input views.
****
Self-supervised Learning of Implicit Shape Representation with Dense Correspondence for Deformable Objects
Baowen Zhang, Jiahe Li, Xiaoming Deng, Yinda Zhang, Cuixia Ma, Hongan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14268-14278
Learning 3D shape representation with dense correspondence for deformable objects is a fundamental problem in computer vision. Existing approaches often need additional annotations of specific semantic domain, e.g., skeleton pose for human body or animals, which require extra annotation effort and suffer from error accumulation, and they are limited to specific domain. In this paper, we propose a novel self-supervised approach to learn neural implicit shape representation for deformable objects, which can represent shapes with a template shape and dense correspondence in 3D. Our method does not require the priors of skeleton and skinning weight, and only requires a collection of shapes represented in signed distance fields. To handle the large deformation, we constrain the learned template shape in the same latent space with the training shapes, design a new formulation of local rigid constraint that enforces rigid transformation in local region and addresses local reflection issue, and present a new hierarchical rigid constraint to reduce the ambiguity due to the joint learning of template shape and correspondence. Extensive experiments show that our model can represent shapes with large deformations. We also show that our shape representation can support two

typical applications, such as texture transfer and shape editing, with competitive performance. The code and models will be publicly released.
****
Perceptual Artifacts Localization for Image Synthesis Tasks
Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, Jianbo Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7579-7590
Recent advancements in deep generative models have facilitated the creation of photo-realistic images across various tasks. However, these generated images often exhibit perceptual artifacts in specific regions, necessitating manual correction. In this study, we present a comprehensive empirical examination of Perceptual Artifacts Localization (PAL) spanning diverse image synthesis endeavors. We introduce a novel dataset comprising 10,168 generated images, each annotated with per-pixel perceptual artifact labels across ten synthesis tasks. A segmentation model, trained on our proposed dataset, effectively localizes artifacts across a range of tasks. Additionally, we illustrate its proficiency in adapting to previously unseen models using minimal training samples. We further propose an innovative zoom-in inpainting pipeline that seamlessly rectifies perceptual artifacts in the generated images. Through our experimental analyses, we elucidate several invaluable downstream applications, such as automated artifact rectification, non-referential image quality evaluation, and abnormal region detection in images. The dataset and code are released here: https://owenzlz.github.io/PAL4VST
****
Narrator: Towards Natural Control of Human-Scene Interaction Generation via Relationship Reasoning
Haibiao Xuan, Xiongzheng Li, Jinsong Zhang, Hongwen Zhang, Yebin Liu, Kun Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22268-22278
Naturally controllable human-scene interaction (HSI) generation has an important role in various fields, such as VR/AR content creation and human-centered AI. However, existing methods are unnatural and unintuitive in their controllability, which heavily limits their application in practice. Therefore, we focus on a challenging task of naturally and controllably generating realistic and diverse HSIs from textual descriptions. From human cognition, the ideal generative model should correctly reason about spatial relationships and interactive actions. To that end, we propose Narrator, a novel relationship reasoning-based generative approach using a conditional variation autoencoder for naturally controllable generation given a 3D scene and a textual description. Also, we model global and local spatial relationships in a 3D scene and a textual description respectively based on the scene graph, and introduce a part-level action mechanism to represent interactions as atomic body part states. In particular, benefiting from our relationship reasoning, we further propose a simple yet effective multi-human generation strategy, which is the first exploration for controllable multi-human scene interaction generation. Our extensive experiments and perceptual studies show that Narrator can controllably generate diverse interactions and significantly outperform existing works.
****
Vision Relation Transformer for Unbiased Scene Graph Generation
Gopika Sudhakaran, Devendra Singh Dhami, Kristian Kersting, Stefan Roth; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21882-21893
Recent years have seen a growing interest in Scene Graph Generation (SGG), a comprehensive visual scene understanding task that aims to predict entity relationships using a relation encoder-decoder pipeline stacked on top of an object encoder-decoder backbone. Unfortunately, current SGG methods suffer from an information loss regarding the entities' local-level cues during the relation encoding process. To mitigate this, we introduce the Vision rElation TransfOrmer (VETO), consisting of a novel local-level entity relation encoder. We further observe that many existing SGG methods claim to be unbiased, but are still biased towards either head or tail classes. To overcome this bias, we introduce a Mutually Exclus

ive ExperT (MEET) learning strategy that captures important relation features wi
thout bias towards head or tail classes. Experimental results on the VG and GQA
datasets demonstrate that VETO + MEET boosts the predictive performance by up to
 47% over the state of the art while being   10x smaller.
****
Scaling Data Generation in Vision-and-Language Navigation
Zun Wang, Jialu Li, Yicong Hong, Yi Wang, Qi Wu, Mohit Bansal, Stephen Gould, Ha
o Tan, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer
 Vision (ICCV), 2023, pp. 12009-12020
Recent research in language-guided visual navigation has demonstrated a signific
ant demand for the diversity of traversable environments and the quantity of sup
ervision for training generalizable agents. To tackle the common data scarcity i
ssue in existing vision-and-language navigation datasets, we propose an effectiv
e paradigm for generating large-scale data for learning, which applies 1200+ pho
to-realistic environments from HM3D and Gibson datasets and synthesizes 4.9 mill
ion instruction trajectory pairs using fully-accessible resources on the web. Im
portantly, we investigate the influence of each component in this paradigm on th
e agent's performance and study how to adequately apply the augmented data to pr
e-train and fine-tune an agent. Thanks to our large-scale dataset, the performan
ce of an existing agent can be pushed up (+11% absolute with regard to previous
SoTA) to a significantly new best of 80% single-run success rate on the R2R test
 split by simple imitation learning. The long-lasting generalization gap between
 navigating in seen and unseen environments is also reduced to less than 1% (ver
sus 8% in the previous best method). Moreover, our paradigm also facilitates dif
ferent models to achieve new state-of-the-art navigation results on CVDN, REVERI
E, and R2R in continuous environments.
****
Better May Not Be Fairer: A Study on Subgroup Discrepancy in Image Classificatio
n
Ming-Chang Chiu, Pin-Yu Chen, Xuezhe Ma; Proceedings of the IEEE/CVF Internation
al Conference on Computer Vision (ICCV), 2023, pp. 4956-4966
In this paper, we provide 20,000 non-trivial human annotations on popular datase
ts as a first step to bridge gap to studying how natural semantic spurious featu
res affect image classification, as prior works often study datasets mixing low-
level features due to limitations in accessing realistic datasets. We investigat
e how natural background colors play a role as spurious features by annotating t
he test sets of CIFAR10 and CIFAR100 into subgroups based on the background colo
r of each image. We name our datasets CIFAR10-B and CIFAR100-B and integrate the
m with CIFAR-Cs. We find that overall human-level accuracy does not guarantee co
nsistent subgroup performances, and the phenomenon remains even on models pre-tr
ained on ImageNet or after data augmentation (DA). To alleviate this issue, we p
ropose FlowAug, a semantic DA that leverages decoupled semantic representations
captured by a pre-trained generative flow. Experimental results show that FlowAu
g achieves more consistent subgroup results than other types of DA methods on CI
FAR10/100 and on CIFAR10/100-C. Additionally, it shows better generalization per
formance. Furthermore, we propose a generic metric, MacroStd, for studying model
 robustness to spurious correlations, where we take a macro average on the weigh
ted standard deviations across different classes. We show MacroStd being more pr
edictive of better performances; per our metric, FlowAug demonstrates improvemen
ts on subgroup discrepancy. Although this metric is proposed to study our curate
d datasets, it applies to all datasets that have subgroups or subclasses. Lastly
, we also show superior out-of-distribution results on CIFAR10.1.
****
3D Implicit Transporter for Temporally Consistent Keypoint Discovery
Chengliang Zhong, Yuhang Zheng, Yupeng Zheng, Hao Zhao, Li Yi, Xiaodong Mu, Ling
 Wang, Pengfei Li, Guyue Zhou, Chao Yang, Xinliang Zhang, Jian Zhao; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3
869-3880
Keypoint-based representation has proven advantageous in various visual and robo
tic tasks. However, the existing 2D and 3D methods for detecting keypoints mainl

y rely on geometric consistency to achieve spatial alignment, neglecting tempora
l consistency. To address this issue, the Transporter method was introduced for
2D data, which reconstructs the target frame from the source frame to incorporat
e both spatial and temporal information. However, the direct application of the
Transporter to 3D point clouds is infeasible due to their structural differences
 from 2D images. Thus, we propose the first 3D version of the Transporter, which
 leverages hybrid 3D representation, cross attention, and implicit reconstructio
n. We apply this new learning system on 3D articulated objects/humans and show t
hat learned keypoints are spatiotemporal consistent. Additionally, we propose a
control policy that utilizes the learned keypoints for 3D object manipulation an
d demonstrate its superior performance. Our codes, data, and models will be made
 publicly available.
****
Adaptive Rotated Convolution for Rotated Object Detection
Yifan Pu, Yiru Wang, Zhuofan Xia, Yizeng Han, Yulin Wang, Weihao Gan, Zidong Wan
g, Shiji Song, Gao Huang; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2023, pp. 6589-6600
Rotated object detection aims to identify and locate objects in images with arbi
trary orientation. In this scenario, the oriented directions of objects vary con
siderably across different images, while multiple orientations of objects exist
within an image. This intrinsic characteristic makes it challenging for standard
 backbone networks to extract high-quality features of these arbitrarily orienta
ted objects. In this paper, we present Adaptive Rotated Convolution (ARC) module
 to handle the aforementioned challenges. In our ARC module, the convolution ker
nels rotate adaptively to extract object features with varying orientations in d
ifferent images, and an efficient conditional computation mechanism is introduce
d to accommodate the large orientation variations of objects within an image. Th
e two designs work seamlessly in rotated object detection problem. Moreover, ARC
 can conveniently serve as a plug-and-play module in various vision backbones to
 boost their representation ability to detect oriented objects accurately. Exper
iments on commonly used benchmarks (DOTA and HRSC2016) demonstrate that equipped
 with our proposed ARC module in the backbone network, the performance of multip
le popular oriented object detectors is significantly improved (e.g. +3.03% mAP
on Rotated RetinaNet and +4.16% on CFA). Combined with the highly competitive me
thod Oriented R-CNN, the proposed approach achieves state-of-the-art performance
 on the DOTA dataset with 81.77% mAP. Code is available at https://github.com/Le
apLabTHU/ARC.
****
Revisit PCA-based Technique for Out-of-Distribution Detection
Xiaoyuan Guan, Zhouwu Liu, Wei-Shi Zheng, Yuren Zhou, Ruixuan Wang; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19
431-19439
Out-of-distribution (OOD) detection is a desired ability to ensure the reliabili
ty and safety of intelligent systems. A scoring function is often designed to me
asure the degree of any new data being an OOD sample. While most designed scorin
g functions are based on a single source of information (e.g., the classifier's
output, logits, or feature vector), recent studies demonstrate that fusion of mu
ltiple sources
 may help better detect OOD data. In this study, after detailed analysis of the
issue in OOD detection by the conventional principal component analysis (PCA), w
e propose fusing a simple regularized PCA-based reconstruction error with other
source of scoring function to further improve OOD detection performance. In part
icular, when combined with a strong energy score-based OOD method, the regulariz
ed reconstruction error helps achieve new state-of-the-art OOD detection results
 on multiple standard benchmarks. The code is available at https://github.com/SY
SU-MIA-GROUP/pca-based-out-of-distribution-detection.
****
Visually-Prompted Language Model for Fine-Grained Scene Graph Generation in an O
pen World
Qifan Yu, Juncheng Li, Yu Wu, Siliang Tang, Wei Ji, Yueting Zhuang; Proceedings

of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21
560-21571
Scene Graph Generation (SGG) aims to extract <subject, predicate, object> relati
onships in images for vision understanding. Although recent works have made stea
dy progress on SGG, they still suffer long-tail distribution that tail-predicate
s are more costly to train and hard to distinguish due to a small amount of anno
tated data compared to frequent predicates. Existing re-balancing strategies try
 to handle it via prior rules but still are confined to pre-defined conditions,
which are not scalable for various models and datasets. In this paper, we propos
e a Cross-modal prediCate boosting (CaCao) framework, where a visually-prompted
language model is learned to generate diverse fine-grained predicates in a low-r
esource way. The proposed CaCao can be applied in a plug-and-play fashion and au
tomatically strengthen existing SGG to tackle the long-tailed problem. Based on
that, we further introduce a novel Entangled cross-modal prompt approach for ope
n-world predicate scene graph generation (Epic), where models can generalize to
unseen predicates in a zero-shot manner. Comprehensive experiments on three benc
hmark datasets show that CaCao consistently boosts the performance of multiple s
cene graph generation models in a model-agnostic way. Moreover, our Epic achieve
s competitive performance on open-world predicate prediction. The data and code
for this paper are publicly available.
****
FishNet: A Large-scale Dataset and Benchmark for Fish Recognition, Detection, an
d Functional Trait Prediction
Faizan Farooq Khan, Xiang Li, Andrew J. Temple, Mohamed Elhoseiny; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 204
96-20506
Aquatic species are essential components of the world's ecosystem, and the prese
rvation of aquatic biodiversity is crucial for maintaining proper ecosystem func
tioning. Unfortunately, increasing anthropogenic pressures such as overfishing,
climate change, and coastal development pose significant threats to aquatic biod
iversity. To address this challenge, it is necessary to design an automatic aqua
tic species monitoring systems that can help researchers and policymakers better
 understand changes in aquatic ecosystems and take appropriate actions to preser
ve biodiversity.
 However, the development of such systems is impeded by a lack of large-scale di
verse aquatic species datasets.
 Existing aquatic species recognition datasets generally have a limited number o
f species, nor do they provide functional trait data, and so have only narrow po
tential for application.
 To address the need for generalized systems that can recognize, locate, and pre
dict a wide array of species and their functional traits, we present FishNet, a
large-scale diverse dataset containing 94,532 meticulously organized images from
 17,357 aquatic species, organized according to aquatic biological taxonomy (ord
er, family, genus, and species). We further build three benchmarks, i.e., fish c
lassification, fish detection, and functional trait prediction, inspired by ecol
ogical research needs, to facilitate the development of aquatic species recognit
ion systems, and promote further research in the field of aquatic ecology. Our F
ishNet dataset has the potential to encourage the development of more accurate a
nd effective tools for the monitoring and protection of aquatic ecosystems, and
hence take effective action toward the conservation of our planet's aquatic biod
iversity. Our dataset and code will be released at https://fishnet-2023.github.i
o/.
****
Dual Learning with Dynamic Knowledge Distillation for Partially Relevant Video R
etrieval
Jianfeng Dong, Minsong Zhang, Zheng Zhang, Xianke Chen, Daizong Liu, Xiaoye Qu,
Xun Wang, Baolong Liu; Proceedings of the IEEE/CVF International Conference on C
omputer Vision (ICCV), 2023, pp. 11302-11312
Almost all previous text-to-video retrieval works assume that videos are pre-tri
mmed with short durations. However, in practice, videos are generally untrimmed

containing much background content. In this work, we investigate the more practical but challenging Partially Relevant Video Retrieval (PRVR) task, which aims to retrieve partially relevant untrimmed videos with the query input. Particularly, we propose to address PRVR from a new perspective, i.e., distilling the generalization knowledge from the large-scale vision-language pre-trained model and transferring it to a task-specific PRVR network. To be specific, we introduce a Dual Learning framework with Dynamic Knowledge Distillation (DL-DKD), which exploits the knowledge of a large vision-language model as the teacher to guide a student model. During the knowledge distillation, an inheritance student branch is devised to absorb the knowledge from the teacher model. Considering that the large model may be of mediocre performance due to the domain gaps, we further develop an exploration student branch to take the benefits of task-specific information. By jointly training the above two branches in a dual-learning way, our model is able to selectively acquire appropriate knowledge from the teacher model while capturing the task-specific property. In addition, a dynamical knowledge distillation strategy is further devised to adjust the effect of each student branch learning during the training. Experiment results demonstrate that our proposed model achieves state-of-the-art performance on ActivityNet and TVR datasets for PRVR.
****
UniVTG: Towards Unified Video-Language Temporal Grounding
Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2794-2804
Video Temporal Grounding (VTG), which aims to ground target clips from videos (such as consecutive intervals or disjoint shots) according to custom language queries (e.g., sentences or words), is key for video browsing on social media. Most methods in this direction develop task-specific models that are trained with type-specific labels, such as moment retrieval (time interval) and highlight detection (worthiness curve), which limits their abilities to generalize to various VTG tasks and labels. In this paper, we propose to Unify the diverse VTG labels and tasks, dubbed UniVTG, along three directions: Firstly, we revisit a wide range of VTG labels and tasks and define a unified formulation. Based on this, we develop data annotation schemes to create scalable pseudo supervision. Secondly, we develop an effective and flexible grounding model capable of addressing each task and making full use of each label. Lastly, thanks to the unified framework, we are able to unlock temporal grounding pretraining from large-scale diverse labels and develop stronger grounding abilities e.g., zero-shot grounding. Extensive experiments on three tasks (moment retrieval, highlight detection and video summarization) across seven datasets (QVHighlights, Charades-STA, TACoS, Ego4D, YouTube Highlights, TV-Sum, and QFVS) demonstrate the effectiveness and flexibility of our proposed framework. The codes are available at https://github.com/showlab/UniVTG.
****
Disposable Transfer Learning for Selective Source Task Unlearning
Seunghee Koh, Hyounguk Shon, Janghyeon Lee, Hyeong Gwon Hong, Junmo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11752-11760
Transfer learning is widely used for training deep neural networks (DNN) for building a powerful representation. Even after the pre-trained model is adapted for the target task, the representation performance of the feature extractor is retained to some extent. As the performance of the pre-trained model can be considered the private property of the owner, it is natural to seek the exclusive right of the generalized performance of the pre-trained weight. To address this issue, we suggest a new paradigm of transfer learning called disposable transfer learning (DTL), which disposes of only the source task without degrading the performance of the target task. To achieve knowledge disposal, we propose a novel loss named Gradient Collision loss (GC loss). GC loss selectively unlearns the source knowledge by leading the gradient vectors of mini-batches in different directions. Whether the model successfully unlearns the source task is measured by piggy

back learning accuracy (PL accuracy). PL accuracy estimates the vulnerability of knowledge leakage by retraining the scrubbed model on a subset of source data or new downstream data. We demonstrate that GC loss is an effective approach to the DTL problem by showing that the model trained with GC loss retains the performance on the target task with a significantly reduced PL accuracy.
****
Grounding 3D Object Affordance from 2D Interactions in Images
Yuhang Yang, Wei Zhai, Hongchen Luo, Yang Cao, Jiebo Luo, Zheng-Jun Zha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10905-10915
Grounding 3D object affordance seeks to locate objects' "action possibilities" regions in the 3D space, which serves as a link between perception and operation for embodied agents. Existing studies primarily focus on connecting visual affordances with geometry structures, e.g., relying on annotations to declare interactive regions of interest on the object and establishing a mapping between the regions and affordances. However, the essence of learning object affordance is to understand how to use it, and the manner that detaches interactions is limited in generalization. Normally, humans possess the ability to perceive object affordances in the physical world through demonstration images or videos. Motivated by this, we introduce a novel task setting: grounding 3D object affordance from 2D interactions in images, which faces the challenge of anticipating affordance through interactions of different sources. To address this problem, we devise a novel Interaction-driven 3D Affordance Grounding Network (IAG), which aligns the region feature of objects from different sources and models the interactive contexts for 3D object affordance grounding. Besides, we collect a Point-Image Affordance Dataset (PIAD) to support the proposed task. Comprehensive experiments on PIAD demonstrate the reliability of the proposed task and the superiority of our method. The project is available at https://github.com/yyvhang/IAGNet.
****
Fast Globally Optimal Surface Normal Estimation from an Affine Correspondence
Levente Hajder, Lajos Lóczi, Daniel Barath; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3390-3401
We present a new solver for estimating a surface normal from a single affine correspondence in two calibrated views. The proposed approach provides a new globally optimal solution for this over-determined problem and proves that it reduces to a linear system that can be solved extremely efficiently. This allows for performing significantly faster than other recent methods, solving the same problem and obtaining the same globally optimal solution. We demonstrate on 15k image pairs from standard benchmarks that the proposed approach leads to the same results as other optimal algorithms while being, on average, five times faster than the fastest alternative. Besides its theoretical value, we demonstrate that such an approach has clear benefits, e.g., in image-based visual localization, due to not requiring a dense point cloud to recover the surface normal. We show on the Cambridge Landmarks dataset that leveraging the proposed surface normal estimation further improves localization accuracy. Matlab and C++ implementations are also published in the supplementary material.
****
Masked Spatio-Temporal Structure Prediction for Self-supervised Learning on Point Cloud Videos
Zhiqiang Shen, Xiaoxiao Sheng, Hehe Fan, Longguang Wang, Yulan Guo, Qiong Liu, Hao Wen, Xi Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16580-16589
Recently, the community has made tremendous progress in developing effective methods for point cloud video understanding that learn from massive amounts of labeled data. However, annotating point cloud videos is usually notoriously expensive. Moreover, training via one or only a few traditional tasks (e.g., classification) may be insufficient to learn subtle details of the spatio-temporal structure existing in point cloud videos. In this paper, we propose a Masked Spatio-Temporal Structure Prediction (MaST-Pre) method to capture the structure of point cloud videos without human annotations. MaST-Pre is based on spatio-temporal point

-tube masking and consists of two self-supervised learning tasks. First, by reco nstructing masked point tubes, our method is able to capture the appearance info rmation of point cloud videos. Second, to learn motion, we propose a temporal ca rdinality difference prediction task that estimates the change in the number of points within a point tube. In this way, MaST-Pre is forced to model the spatial and temporal structure in point cloud videos. Extensive experiments on MSRActio n-3D, NTU-RGBD, NvGesture, and SHREC'17 demonstrate the effectiveness of the pro posed method.
****
Frequency-aware GAN for Adversarial Manipulation Generation
Peifei Zhu, Genki Osada, Hirokatsu Kataoka, Tsubasa Takahashi; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4315-43 24
Image manipulation techniques have drawn growing concerns as manipulated images might cause morality and security problems. Various methods have been proposed t o detect manipulations and achieved promising performance. However, these method s might be vulnerable to adversarial attacks. In this work, we design an Adversa rial Manipulation Generation (AMG) task to explore the vulnerability of image ma nipulation detectors. We first propose an optimal loss function and extend exist ing attacks to generate adversarial examples. We observe that existing spatial a ttacks cause large degradation in image quality and find the loss of high-freque ncy detailed components might be its major reason. Inspired by this observation, we propose a novel adversarial attack that incorporates both spatial and freque ncy features into the GAN architecture to generate adversarial examples. We furt her design an encoder-decoder architecture with skip connections of high-frequen cy components to preserve fine details. We evaluated our method on three image m anipulation detectors (FCN, ManTra-Net and MVSS-Net) with three benchmark datase ts (DEFACTO, CASIAv2 and COVER). Experiments show that our method generates adve rsarial examples significantly fast (0.01s per image), preserves better image qu ality (PSNR 30% higher than spatial attacks) and achieves a high attack success rate. We also observe that the examples generated by AMG can fool both classific ation and segmentation models, which indicates better transferability among diff erent tasks.
****
DreamPose: Fashion Video Synthesis with Stable Diffusion
Johanna Karras, Aleksander Holynski, Ting-Chun Wang, Ira Kemelmacher-Shlizerman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22680-22690
We present DreamPose, a diffusion-based method for generating animated fashion v ideos from still images. Given an image and a sequence of human body poses, our method synthesizes a video containing both human and fabric motion. To achieve t his, we transform a pretrained text-to-image model (Stable Diffusion) into a pos e-and-image guided video synthesis model, using a novel finetuning strategy, a s et of architectural changes to support the added conditioning signals, and techn iques to encourage temporal consistency. We fine-tune on a collection of fashion videos from the UBC Fashion dataset. We evaluate our method on a variety of clo thing styles and poses, and demonstrate that our method produces state-of-the-ar t results on fashion video animation. Video results are available at www.grail.c s.washington.edu/projects/dreampose.
****
Tube-Link: A Flexible Cross Tube Framework for Universal Video Segmentation
Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, Chen Ch ange Loy; Proceedings of the IEEE/CVF International Conference on Computer Visio n (ICCV), 2023, pp. 13923-13933
Video segmentation aims to segment and track every pixel in diverse scenarios ac curately. In this paper, we present Tube-Link, a versatile framework that addres ses multiple core tasks of video segmentation with a unified architecture. Our f ramework is a near-online approach that takes a short subclip as input and outpu ts the corresponding spatial-temporal tube masks. To enhance the modeling of cro ss-tube relationships, we propose an effective way to perform tube-level linking

via attention along the queries. In addition, we introduce temporal contrastive learning to instance-wise discriminative features for tube-level association. Our approach offers flexibility and efficiency for both short and long video inputs, as the length of each subclip can be varied according to the needs of datasets or scenarios. Tube-Link outperforms existing specialized architectures by a significant margin on five video segmentation datasets. Specifically, it achieves almost 13% relative improvements on VIPSeg and 4% improvements on KITTI-STEP over the strong baseline Video K-Net. When using a ResNet50 backbone on Youtube-VIS-2019 and 2021, Tube-Link boosts IDOL by 3% and 4%, respectively. Code is available at https://github.com/lxtGH/Tube-Link.
****
Hybrid Spectral Denoising Transformer with Guided Attention
Zeqiang Lai, Chenggang Yan, Ying Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13065-13075
In this paper, we present a Hybrid Spectral Denoising Transformer (HSDT) for hyperspectral image denoising. Challenges in adapting transformer for HSI arise from the capabilities to tackle existing limitations of CNN-based methods in capturing the global and local spatial-spectral correlations while maintaining efficiency and flexibility. To address these issues, we introduce a hybrid approach that combines the advantages of both models with a Spatial-Spectral Separable Convolution (S3Conv), Guided Spectral Self-Attention (GSSA), and Self-Modulated Feed-Forward Network (SM-FFN). Our S3Conv works as a lightweight alternative to 3D convolution, which extracts more spatial-spectral correlated features while keeping the flexibility to tackle HSIs with an arbitrary number of bands. These features are then adaptively processed by GSSA which performs 3D self-attention across the spectral bands, guided by a set of learnable queries that encode the spectral signatures. This not only enriches our model with powerful capabilities for identifying global spectral correlations but also maintains linear complexity. Moreover, our SM-FFN proposes the self-modulation that intensifies the activations of more informative regions, which further strengthens the aggregated features. Extensive experiments are conducted on various datasets under both simulated and real-world noise, and it shows that our HSDT significantly outperforms the existing state-of-the-art methods while maintaining low computational overhead. Code is at https://github.com/Zeqiang-Lai/HSDT.
****
HiVLP: Hierarchical Interactive Video-Language Pre-Training
Bin Shao, Jianzhuang Liu, Renjing Pei, Songcen Xu, Peng Dai, Juwei Lu, Weimian Li, Youliang Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13756-13766
Video-Language Pre-training (VLP) has become one of the most popular research topics in deep learning. However, compared to image-language pre-training, VLP has lagged far behind due to the lack of large amounts of video-text pairs. In this work, we train a VLP model with a hybrid of image-text and video-text pairs, which significantly outperforms pre-training with only the video-text pairs. Besides, existing methods usually model the cross-modal interaction using cross-attention between single-scale visual tokens and textual tokens. These visual features are either of low resolutions lacking fine-grained information, or of high resolutions without high-level semantics. To address the issue, we propose Hierarchical interactive Video-Language Pre-training (HiVLP) that efficiently uses a hierarchical visual feature group for multi-modal cross-attention during pre-training. In the hierarchical framework, low-resolution features are learned with focus on more global high-level semantic information, while high-resolution features carry fine-grained details. As a result, HiVLP has the ability to effectively learn both the global and fine-grained representations to achieve better alignment between video and text inputs. Furthermore, we design a hierarchical multi-scale vision contrastive loss for self-supervised learning to boost the interaction between them. Experimental results show that HiVLP establishes new state-of-the-art results in three downstream tasks, text-video retrieval, video-text retrieval, and video captioning.
****

Learning Concordant Attention via Target-aware Alignment for Visible-Infrared Person Re-identification
Jianbing Wu, Hong Liu, Yuxin Su, Wei Shi, Hao Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11122-11131
Owing to the large distribution gap between the heterogeneous data in Visible-Infrared Person Re-identification (VI Re-ID), we point out that existing paradigms often suffer from the inter-modal semantic misalignment issue and thus fail to align and compare local details properly. In this paper, we present Concordant Attention Learning (CAL), a novel framework that learns semantic-aligned representations for VI Re-ID. Specifically, we design the Target-aware Concordant Alignment paradigm, which allows target-aware attention adaptation when aligning heterogeneous samples (i.e., adaptive attention adjustment according to the target image being aligned). This is achieved by exploiting the discriminative clues from the modality counterpart and designing effective modality-agnostic correspondence searching strategies. To ensure semantic concordance during the cross-modal retrieval stage, we further propose MatchDistill, which matches the attention patterns across modalities and learns their underlying semantic correlations by bipartite-graph-based similarity modeling and cross-modal knowledge exchange. Extensive experiments on VI Re-ID benchmark datasets demonstrate the effectiveness and superiority of the proposed CAL.
****
PhysDiff: Physics-Guided Human Motion Diffusion Model
Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, Jan Kautz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16010-16021
Denoising diffusion models hold great promise for generating diverse and realistic human motions. However, existing motion diffusion models largely disregard the laws of physics in the diffusion process and often generate physically-implausible motions with pronounced artifacts such as floating, foot sliding, and ground penetration. This seriously impacts the quality of generated motions and limits their real-world application. To address this issue, we present a novel physics-guided motion diffusion model (PhysDiff), which incorporates physical constraints into the diffusion process. Specifically, we propose a physics-based motion projection module that uses motion imitation in a physics simulator to project the denoised motion of a diffusion step to a physically-plausible motion. The projected motion is further used in the next diffusion step to guide the denoising diffusion process. Intuitively, the use of physics in our model iteratively pulls the motion toward a physically-plausible space, which cannot be achieved by simple post-processing. Experiments on large-scale human motion datasets show that our approach achieves state-of-the-art motion quality and improves physical plausibility drastically (>78% for all datasets).
****
Masked Motion Predictors are Strong 3D Action Representation Learners
Yunyao Mao, Jiajun Deng, Wengang Zhou, Yao Fang, Wanli Ouyang, Houqiang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10181-10191
In 3D human action recognition, limited supervised data makes it challenging to fully tap into the modeling potential of powerful networks such as transformers. As a result, researchers have been actively investigating effective self-supervised pre-training strategies. In this work, we show that instead of following the prevalent pretext task to perform masked self-component reconstruction in human joints, explicit contextual motion modeling is key to the success of learning effective feature representation for 3D action recognition. Formally, we propose the Masked Motion Prediction (MAMP) framework. To be specific, the proposed MAMP takes as input the masked spatio-temporal skeleton sequence and predicts the corresponding temporal motion of the masked human joints. Considering the high temporal redundancy of the skeleton sequence, in our MAMP, the motion information also acts as an empirical semantic richness prior that guide the masking process, promoting better attention to semantically rich temporal regions. Extensive experiments on NTU-60, NTU-120, and PKU-MMD datasets show that the proposed MAMP p

re-training substantially improves the performance of the adopted vanilla transf
ormer, achieving state-of-the-art results without bells and whistles. The source
 code of our MAMP is available at https://github.com/maoyunyao/MAMP.
****
Template-guided Hierarchical Feature Restoration for Anomaly Detection
Hewei Guo, Liping Ren, Jingjing Fu, Yuwang Wang, Zhizheng Zhang, Cuiling Lan, Ha
oqian Wang, Xinwen Hou; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2023, pp. 6447-6458
Targeting for detecting anomalies of various sizes for complicated normal patter
ns, we propose a Template-guided Hierarchical Feature Restoration method, which
introduces two key techniques, bottleneck compression and template-guided compen
sation, for anomaly-free feature restoration. Specially, our framework compresse
s hierarchical features of an image by bottleneck structure to preserve the most
 crucial features shared among normal samples. We design template-guided compens
ation to restore the distorted features towards anomaly-free features. Particula
rly, we choose the most similar normal sample as the template and leverage hiera
rchical features from the template to compensate the distorted features. The bot
tleneck could partially filter out anomaly features, while the compensation furt
her converts the reminding anomaly features towards normal with template guidanc
e. Finally, anomalies are detected in terms of the cosine distance between the p
re-trained features of an inference image and the corresponding restored anomaly
-free features. Experimental results demonstrate the effectiveness of our approa
ch, which achieves the state-of-the-art performance on the MVTec LOCO AD dataset
.
****
SwiftFormer: Efficient Additive Attention for Transformer-based Real-time Mobile
 Vision Applications
Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang
, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2023, pp. 17425-17436
Self-attention has become a defacto choice for capturing global context in vario
us vision applications. However, its quadratic computational complexity with res
pect to image resolution limits its use in real-time applications, especially fo
r deployment on resource-constrained mobile devices. Although hybrid approaches
have been proposed to combine the advantages of convolutions and self-attention
for a better speed-accuracy trade-off, the expensive matrix multiplication opera
tions in self-attention remain a bottleneck. In this work, we introduce a novel
efficient additive attention mechanism that effectively replaces the quadratic m
atrix multiplication operations with linear element-wise multiplications. Our de
sign shows that the key-value interaction can be replaced with a linear layer wi
thout sacrificing any accuracy. Unlike previous state-of-the-art methods, our ef
ficient formulation of self-attention enables its usage at all stages of the net
work. Using our proposed efficient additive attention, we build a series of mode
ls called "SwiftFormer" which achieves state-of-the-art performance in terms of
both accuracy and mobile inference speed. Our small variant achieves 78.5% top-1
 ImageNet-1K accuracy with only 0.8 ms latency on iPhone 14, which is more accur
ate and 2x faster compared to MobileViT-v2. Our code and models: https://tinyurl
.com/5ft8v46w
****
UpCycling: Semi-supervised 3D Object Detection without Sharing Raw-level Unlabel
ed Scenes
Sunwook Hwang, Youngseok Kim, Seongwon Kim, Saewoong Bahk, Hyung-Sin Kim; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 23351-23361
Semi-supervised Learning (SSL) has received increasing attention in autonomous d
riving to reduce the enormous burden of 3D annotation. In this paper, we propose
 UpCycling, a novel SSL framework for 3D object detection with zero additional r
aw-level point cloud: learning from unlabeled de-identified intermediate feature
s (i.e., "smashed" data) to preserve privacy. Since these intermediate features
are naturally produced by the inference pipeline, no additional computation is r

equired on autonomous vehicles. However, generating effective consistency loss for unlabeled feature-level scene turns out to be a critical challenge. The latest SSL frameworks for 3D object detection that enforce consistency regularization between different augmentations of an unlabeled raw-point scene become detrimental when applied to intermediate features. To solve the problem, we introduce a novel combination of hybrid pseudo labels and feature-level Ground Truth sampling (F-GT), which safely augments unlabeled multi-type 3D scene features and provides high-quality supervision. We implement UpCycling on two representative 3D object detection models: SECOND-IoU and PV-RCNN. Experiments on widely-used datasets (Waymo, KITTI, and Lyft) verify that UpCycling outperforms other augmentation methods applied at the feature level. In addition, while preserving privacy, UpCycling performs better or comparably to the state-of-the-art methods that utilize raw-level unlabeled data in both domain adaptation and partial-label scenarios.
****
RIGID: Recurrent GAN Inversion and Editing of Real Face Videos
Yangyang Xu, Shengfeng He, Kwan-Yee K. Wong, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13691-13701
GAN inversion is indispensable for applying the powerful editability of GAN to real images. However, existing methods invert video frames individually often leading to undesired inconsistent results over time. In this paper, we propose a unified recurrent framework, named Recurrent vIdeo GAN Inversion and eDiting (RIGID), to explicitly and simultaneously enforce temporally coherent GAN inversion and facial editing of real videos. Our approach models the temporal relations between current and previous frames from three aspects. To enable a faithful real video reconstruction, we first maximize the inversion fidelity and consistency by learning a temporal compensated latent code. Second, we observe incoherent noises lie in the high-frequency domain that can be disentangled from the latent space. Third, to remove the inconsistency after attribute manipulation, we propose an in-between frame composition constraint such that the arbitrary frame must be a direct composite of its neighboring frames. Our unified framework learns the inherent coherence between input frames in an end-to-end manner, and therefore it is agnostic to a specific attribute and can be applied to arbitrary editing of the same video without re-training. Extensive experiments demonstrate that RIGID outperforms state-of-the-art methods qualitatively and quantitatively in both inversion and editing tasks. The deliverables can be found in https://cnnlstm.github.io/RIGID.
****
PourIt!: Weakly-Supervised Liquid Perception from a Single Image for Visual Closed-Loop Robotic Pouring
Haitao Lin, Yanwei Fu, Xiangyang Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 241-251
Liquid perception is critical for robotic pouring tasks. It usually requires the robust visual detection of flowing liquid. However, while recent works have shown promising results in liquid perception, they typically require labeled data for model training, a process that is both time-consuming and reliant on human labor. To this end, this paper proposes a simple yet effective framework PourIt!, to serve as a tool for robotic pouring tasks. We design a simple data collection pipeline that only needs image-level labels to reduce the reliance on tedious pixel-wise annotations. Then, a binary classification model is trained to generate Class Activation Map (CAM) that focuses on the visual difference between these two kinds of collected data, i.e., the existence of liquid drop or not. We also devise a feature contrast strategy to improve the quality of the CAM, thus entirely and tightly covering the actual liquid regions. Then, the container pose is further utilized to facilitate the 3D point cloud recovery of the detected liquid region. Finally, the liquid-to-container distance is calculated for visual closed-loop control of the physical robot. To validate the effectiveness of our proposed method, we also contribute a novel dataset for our task and name it PourIt! dataset. Extensive results on this dataset and physical Franka robot have shown the utility and effectiveness of our method in the robotic pouring tasks. Our

dataset, code and pre-trained models will be available on the project page.
****

CSDA: Learning Category-Scale Joint Feature for Domain Adaptive Object Detection
Changlong Gao, Chengxu Liu, Yujie Dun, Xueming Qian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11421-11430
Domain Adaptive Object Detection (DAOD) aims to improve the detection performance of target domains by minimizing the feature distribution between the source and target domain. Recent approaches usually align such distributions in terms of categories through adversarial learning and some progress has been made. However, when objects are non-uniformly distributed at different scales, such category-level alignment causes imbalanced object feature learning, refer as the inconsistency of category alignment at different scales. For better category-level feature alignment, we propose a novel DAOD framework of joint category and scale information, dubbed CSDA, such a design enables effective object learning for different scales. Specifically, our framework is implemented by two closely-related modules: 1) SGFF (Scale-Guided Feature Fusion) fuses the category representations of different domains to learn category-specific features, where the features are aligned by discriminators at three scales. 2) SAFE (Scale-Auxiliary Feature Enhancement) encodes scale coordinates into a group of tokens and enhances the representation of category-specific features at different scales by self-attention. Based on the anchor-based Faster-RCNN and anchor-free FCOS detectors, experiments show that our method achieves state-of-the-art results on three DAOD benchmarks.
****

A Latent Space of Stochastic Diffusion Models for Zero-Shot Image Editing and Guidance
Chen Henry Wu, Fernando De la Torre; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7378-7387
Diffusion models generate images by iterative denoising. Recent work has shown that by making the denoising process deterministic, one can encode real images into latent codes of the same size, which can be used for image editing. This paper explores the possibility of defining a latent space even when the denoising process remains stochastic. Recall that, in stochastic diffusion models, Gaussian noises are added in each denoising step, and we can concatenate all the noises to form a latent code. This results in a latent space of much higher dimensionality than the original image. We demonstrate that this latent space of stochastic diffusion models can be used in the same way as that of deterministic diffusion models in two applications. First, we propose CycleDiffusion, a method for zero-shot and unpaired image editing using stochastic diffusion models, which improves the performance over its deterministic counterpart. Second, we demonstrate unified, plug-and-play guidance in the latent spaces of deterministic and stochastic diffusion models.
****

Single Image Defocus Deblurring via Implicit Neural Inverse Kernels
Yuhui Quan, Xin Yao, Hui Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12600-12610
Single image defocus deblurring (SIDD) is a challenging task due to the spatially-varying nature of defocus blur, characterized by per-pixel point spread functions (PSFs). Existing deep-learning-based methods for SIDD are limited by either over-fitting due to the lack of model constraints or under-parametrization that restricts their applicability to real-world images. To address the limitations, this paper proposes an interpretable approach that explicitly predicts inverse kernels with structural regularization. Motivated by the observation that defocus PSFs within an image often have similar shapes but different sizes, we represent the inverse kernels linearly over a multi-scale dictionary parameterized by implicit neural representations. We predict the corresponding representation coefficients via a duplex scale-recurrent neural network that jointly performs fine-to-coarse and coarse-to-fine estimations. Extensive experiments demonstrate that our approach achieves excellent performance using a lightweight model.
****

Open Set Video HOI detection from Action-Centric Chain-of-Look Prompting
Nan Xi, Jingjing Meng, Junsong Yuan; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 3079-3089
Human-Object Interaction (HOI) detection is essential for understanding and mode
ling real-world events. Existing works on HOI detection mainly focus on static i
mages and a closed setting, where all HOI classes are provided in the training s
et. In comparison, detecting HOIs in videos in open set scenarios is more challe
nging. First, under open set circumstances, HOI detectors are expected to hold s
trong generalizability to recognize unseen HOIs not included in the training dat
a. Second, accurately capturing temporal contextual information from videos is d
ifficult, but it is crucial for detecting temporal-related actions
such as open, close, pull, push. To this end, we propose ACoLP, a model of Acti
on-centric Chain-of-Look Prompting for open set video HOI detection. ACoLP regar
ds actions as the carrier of semantics in videos, which captures the essential s
emantic information across frames. To make the model generalizable on unseen cla
sses, inspired by the chain-of-thought prompting in natural language processing,
we introduce the chain-of-look prompting scheme that decomposes prompt generati
on from large-scale vision-language model into a series of intermediate visual r
easoning steps. Consequently, our model captures
complex visual reasoning processes underlying the HOI events in videos, providi
ng essential guidance for detecting unseen classes. Extensive experiments on two
video HOI datasets, VidHOI and CAD120, demonstrate that ACoLP achieves competit
ive performance compared with the state-of-the-art methods in the conventional c
losed setting, and outperforms existing methods by a large margin in the open se
t setting. Our code is avaliable at https://github.
com/southnx/ACoLP.
****
Robust Mixture-of-Expert Training for Convolutional Neural Networks
Yihua Zhang, Ruisi Cai, Tianlong Chen, Guanhua Zhang, Huan Zhang, Pin-Yu Chen, S
hiyu Chang, Zhangyang Wang, Sijia Liu; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2023, pp. 90-101
Sparsely-gated Mixture of Expert (MoE), an emerging deep model architecture, has
demonstrated a great promise to enable high-accuracy and ultra-efficient model
inference. Despite the growing popularity of MoE, little work investigated its p
otential to advance convolutional neural networks (CNNs), especially in the plan
e of adversarial robustness. Since the lack of robustness has become one of the
main hurdles for CNNs, in this paper we ask: How to adversarially robustify a CN
N-based MoE model? Can we robustly train it like an ordinary CNN model? Our pilo
t study shows that the conventional adversarial training (AT) mechanism (develop
ed for vanilla CNNs) no longer remains effective to robustify an MoE-CNN. To bet
ter understand this phenomenon, we dissect the robustness of an MoE-CNN into two
dimensions: Robustness of routers (i.e., gating functions to select data-specif
ic experts) and robustness of experts (i.e., the router-guided pathways defined
by the subnetworks of the backbone CNN). Our analyses show that routers and expe
rts are hard to adapt to each other in the vanilla AT. Thus, we propose a new ro
uter-expert alternating Adversarial training framework for MoE, termed AdvMoE. T
he effectiveness of our proposal is justified across 4 commonly-used CNN model a
rchitectures over 4 benchmark datasets. We find that AdvMoE achieves 1%   4% adv
ersarial robustness improvement over the original dense CNN, and enjoys the effi
ciency merit of sparsity-gated MoE, leading to more than 50% inference cost redu
ction.
****
AvatarCraft: Transforming Text into Neural Human Avatars with Parameterized Shap
e and Pose Control
Ruixiang Jiang, Can Wang, Jingbo Zhang, Menglei Chai, Mingming He, Dongdong Chen
, Jing Liao; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2023, pp. 14371-14382
Neural implicit fields are powerful for representing 3D scenes and generating hi
gh-quality novel views, but it remains challenging to use such implicit represen
tations for creating a 3D human avatar with a specific identity and artistic sty

le that can be easily animated. Our proposed method, AvatarCraft, addresses this challenge by using diffusion models to guide the learning of geometry and texture for a neural avatar based on a single text prompt. We carefully design the optimization framework of neural implicit fields, including a coarse-to-fine multi-bounding box training strategy, shape regularization, and diffusion-based constraints, to produce high-quality geometry and texture. Additionally, we make the human avatar animatable by deforming the neural implicit field with an explicit warping field that maps the target human mesh to a template human mesh, both represented using parametric human models. This simplifies animation and reshaping of the generated avatar by controlling pose and shape parameters. Extensive experiments on various text descriptions show that AvatarCraft is effective and robust in creating human avatars and rendering novel views, poses, and shapes. Our project page is: https://avatar-craft.github.io/.
****
s-Adaptive Decoupled Prototype for Few-Shot Object Detection
Jinhao Du, Shan Zhang, Qiang Chen, Haifeng Le, Yanpeng Sun, Yao Ni, Jian Wang, Bin He, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18950-18960
Meta-learning-based few-shot detectors use one K-average-pooled prototype (averaging along K-shot dimension) in both Region Proposal Network (RPN) and Detection head (DH) for query detection. Such plain operation would harm the FSOD performance in two aspects: 1) the poor quality of the prototype, and 2) the equivocal guidance due to the contradictions between RPN and DH. In this paper, we look closely into those critical issues and propose the s-Adaptive Decoupled Prototype (s-ADP) as a solution. To generate the high-quality prototype, we prioritize salient representations and deemphasize trivial variations by accessing both angle distance and magnitude dispersion (s) across K-support samples. To provide precise information for the query image, the prototype is decoupled into task-specific ones, which provide tailored guidance for 'where to look' and 'what to look for', respectively.
 Beyond that, we find our s-ADP can gradually strengthen the generalization power of encoding network during meta-training. So it can robustly deal with intra-class variations and a simple K- average pooling is enough to generate a high-quality prototype at meta-testing. We provide theoretical analysis to support its rationality. Extensive experiments on Pascal VOC, MS-COCO and FSOD datasets demonstrate that the proposed method achieves new state-of-the-art performance. Notably, our method surpasses the baseline model by a large margin - up to around 5.0% AP50 and 8.0% AP75 on novel classes.
****
Why Is Prompt Tuning for Vision-Language Models Robust to Noisy Labels?
Cheng-En Wu, Yu Tian, Haichao Yu, Heng Wang, Pedro Morgado, Yu Hen Hu, Linjie Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15488-15497
Vision-language models such as CLIP learn a generic text-image embedding from large-scale training data. A vision-language model can be adapted to a new classification task through few-shot prompt tuning. We find that such prompt tuning process is highly robust to label noises. This intrigues us to study the key reasons contributing to the robustness of the prompt tuning paradigm. We conducted extensive experiments to explore this property and find the key factors are: 1. the fixed classname tokens provide a strong regularization to the optimization of the model, reducing gradients induced by the noisy samples; 2. the powerful pre-trained image-text embedding that is learned from diverse and generic web data provides strong prior knowledge for image classification. Further, we demonstrate that noisy zero-shot predictions from CLIP can be used to tune its own prompt, significantly enhancing prediction accuracy in the unsupervised setting.
****
Unified Pre-Training with Pseudo Texts for Text-To-Image Person Re-Identification
Zhiyin Shao, Xinyu Zhang, Changxing Ding, Jian Wang, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11

174-11184
The pre-training task is indispensable for the text-to-image person re-identific
ation (T2I-ReID) task. However, there are two underlying inconsistencies between
 these two tasks that may impact the performance: i) Data inconsistency. A large
 domain gap exists between the generic images/texts used in public pre-trained m
odels and the specific person data in the T2I-ReID task. This gap is especially
severe for texts, as general textual data are usually unable to describe specifi
c people in fine-grained detail. ii) Training inconsistency. The processes of pr
e-training of images and texts are independent, despite cross-modality learning
being critical to T2I-ReID. To address the above issues, we present a new unifie
d pre-training pipeline (UniPT) designed specifically for the T2I-ReID task. We
first build a large-scale text-labeled person dataset "LUPerson-T", in which pse
udo-textual descriptions of images are automatically generated by the CLIP parad
igm using a divide-conquer-combine strategy. Benefiting from this dataset, we th
en utilize a simple vision-and-language pre-training framework to explicitly ali
gn the feature space of the image and text modalities during pre-training. In th
is way, the pre-training task and the T2I-ReID task are made consistent with eac
h other on both data and training levels. Without the need for any bells and whi
stles, our UniPT achieves competitive Rank-1 accuracy of, i.e., 68.50%, 60.09%,
and 51.85% on CUHK-PEDES, ICFG-PEDES and RSTPReid, respectively. Both the LUPers
on-T dataset and code are available at https://github.com/ZhiyinShao-H/UniPT.
****
Semantics Meets Temporal Correspondence: Self-supervised Object-centric Learning
 in Videos
Rui Qian, Shuangrui Ding, Xian Liu, Dahua Lin; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2023, pp. 16675-16687
Self-supervised methods have shown remarkable progress in learning high-level se
mantics and low-level temporal correspondence. Building on these results, we tak
e one step further and explore the possibility of integrating these two features
 to enhance object-centric representations. Our preliminary experiments indicate
 that query slot attention can extract different semantic components from the RG
B feature map, while random sampling based slot attention can exploit temporal c
orrespondence cues between frames to assist instance identification. Motivated b
y this, we propose a novel semantic-aware masked slot attention on top of the fu
sed semantic features and correspondence maps. It comprises two slot attention s
tages with a set of shared learnable Gaussian distributions. In the first stage,
 we use the mean vectors as slot initialization to decompose potential semantics
 and generate semantic segmentation masks through iterative attention. In the se
cond stage, for each semantics, we randomly sample slots from the corresponding
Gaussian distribution and perform masked feature aggregation within the semantic
 area to exploit temporal correspondence patterns for instance identification. W
e adopt semantic- and instance-level temporal consistency as self-supervision to
 encourage temporally coherent object-centric representations. Our model effecti
vely identifies multiple object instances with semantic structure, reaching prom
ising results on unsupervised video object discovery. Furthermore, we achieve st
ate-of-the-art performance on dense label propagation tasks, demonstrating the p
otential for object-centric analysis.
****
UniTR: A Unified and Efficient Multi-Modal Transformer for Bird's-Eye-View Repre
sentation
Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, Liwe
i Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2023, pp. 6792-6802
Jointly processing information from multiple sensors is crucial to achieving acc
urate and robust perception for reliable autonomous driving systems. However, cu
rrent 3D perception research follows a modality-specific paradigm, leading to ad
ditional computation overheads and inefficient collaboration between different s
ensor data. In this paper, we present an efficient multi-modal backbone for outd
oor 3D perception named UniTR, which processes a variety of modalities with unif
ied modeling and shared parameters. Unlike previous works, UniTR introduces a mo

dality-agnostic transformer encoder to handle these view-discrepant sensor data for parallel modal-wise representation learning and automatic cross-modal interaction without additional fusion steps. More importantly, to make full use of these complementary sensor types, we present a novel multi-modal integration strategy by both considering semantic-abundant 2D perspective and geometry-aware 3D sparse neighborhood relations. UniTR is also a fundamentally task-agnostic backbone that naturally supports different 3D perception tasks. It sets a new state-of-the-art performance on the nuScenes benchmark, achieving +1.1 NDS higher for 3D object detection and +12.0 higher mIoU for BEV map segmentation with lower inference latency. Code will be available at https://github.com/Haiyang-W/UniTR.
****

Traj-MAE: Masked Autoencoders for Trajectory Prediction

Hao Chen, Jiaze Wang, Kun Shao, Furui Liu, Jianye Hao, Chenyong Guan, Guangyong Chen, Pheng-Ann Heng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8351-8362

Trajectory prediction has been a crucial task in building a reliable autonomous driving system by anticipating possible dangers. One key issue is to generate consistent trajectory predictions without colliding. To overcome the challenge, we propose an efficient masked autoencoder for trajectory prediction (Traj-MAE) that better represents the complicated behaviors of agents in the driving environment. Specifically, our Traj-MAE employs diverse masking strategies to pre-train the trajectory encoder and map encoder, allowing for the capture of social and temporal information among agents while leveraging the effect of environment from multiple granularities. To address the catastrophic forgetting problem that arises when pre-training the network with multiple masking strategies, we introduce a continual pre-training framework, which can help Traj-MAE learn valuable and diverse information from various strategies efficiently. Our experimental results in both multi-agent and single-agent settings demonstrate that Traj-MAE achieves competitive results with state-of-the-art methods and significantly outperforms our baseline model. Project page: https://jiazewang.com/projects/trajmae.html.
****

First Session Adaptation: A Strong Replay-Free Baseline for Class-Incremental Learning

Aristeidis Panos, Yuriko Kobe, Daniel Olmeda Reino, Rahaf Aljundi, Richard E. Turner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18820-18830

In Class-Incremental Learning (CIL) an image classification system is exposed to new classes in each learning session and must be updated incrementally. Methods approaching this problem have updated both the classification head and the feature extractor body at each session of CIL. In this work, we develop a baseline method, First Session Adaptation (FSA), that sheds light on the efficacy of existing CIL approaches, and allows us to assess the relative performance contributions from head and body adaption. FSA adapts a pre-trained neural network body only on the first learning session and fixes it thereafter; a head based on linear discriminant analysis (LDA), is then placed on top of the adapted body, allowing exact updates through CIL. FSA is replay-free i.e. it does not memorize examples from previous sessions of continual learning. To empirically motivate FSA, we first consider a diverse selection of 22 image-classification datasets, evaluating different heads and body adaptation techniques in high/low-shot offline settings. We find that the LDA head performs well and supports CIL out-of-the-box. We also find that Featurewise Layer Modulation (FiLM) adapters are highly effective in the few-shot setting, and full-body adaption in the high-shot setting. Second, we empirically investigate various CIL settings including high-shot CIL and few-shot CIL, including settings that have previously been used in the literature. We show that FSA significantly improves over the state-of-the-art in 15 of the 16 settings considered. FSA with FiLM adapters is especially performant in the few-shot setting. These results indicate that current approaches to continuous body adaptation are not working as expected. Finally, we propose a measure that can be applied to a set of unlabelled inputs which is predictive of the benefits

of body adaptation.
****
Ada3D : Exploiting the Spatial Redundancy with Adaptive Inference for Efficient 3D Object Detection
Tianchen Zhao, Xuefei Ning, Ke Hong, Zhongyuan Qiu, Pu Lu, Yali Zhao, Linfeng Zhang, Lipu Zhou, Guohao Dai, Huazhong Yang, Yu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17728-17738
Voxel-based methods have achieved state-of-the-art performance for 3D object detection in autonomous driving. However, their significant computational and memory costs pose a challenge for their application to resource-constrained vehicles. One reason for this high resource consumption is the presence of a large number of redundant background points in Lidar point clouds, resulting in spatial redundancy in both 3D voxel and dense BEV map representations. To address this issue, we propose an adaptive inference framework called Ada3D, which focuses on exploiting the input-level spatial redundancy. Ada3D adaptively filters the redundant input, guided by a lightweight importance predictor and the unique properties of the Lidar point cloud. Additionally, we utilize the BEV features' intrinsic sparsity by introducing the Sparsity Preserving Batch Normalization. With Ada3D, we achieve 40% reduction for 3D voxels and decrease the density of 2D BEV feature maps from 100% to 20% without sacrificing accuracy. Ada3D reduces the model computational and memory cost by 5x, and achieves 1.52x/1.45x end-to-end GPU latency and 1.5x/4.5x GPU peak memory optimization for the 3D and 2D backbone respectively.
****
R3D3: Dense 3D Reconstruction of Dynamic Scenes from Multiple Cameras
Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, Fisher Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3216-3226
Dense 3D reconstruction and ego-motion estimation are key challenges in autonomous driving and robotics. Compared to the complex, multi-modal systems deployed today, multi-camera systems provide a simpler, low-cost alternative. However, camera-based 3D reconstruction of complex dynamic scenes has proven extremely difficult, as existing solutions often produce incomplete or incoherent results. We propose R3D3, a multi-camera system for dense 3D reconstruction and ego-motion estimation. Our approach iterates between geometric estimation that exploits spatial-temporal information from multiple cameras, and monocular depth refinement. We integrate multi-camera feature correlation and dense bundle adjustment operators that yield robust geometric depth and pose estimates. To improve reconstruction where geometric depth is unreliable, e.g. for moving objects or low-textured regions, we introduce learnable scene priors via a depth refinement network. We show that this design enables a dense, consistent 3D reconstruction of challenging, dynamic outdoor environments. Consequently, we achieve state-of-the-art dense depth prediction on the DDAD and NuScenes benchmarks.
****
UniFusion: Unified Multi-View Fusion Transformer for Spatial-Temporal Representation in Bird's-Eye-View
Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, Xi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8690-8699
Bird's eye view (BEV) representation is a new perception formulation for autonomous driving, which is based on spatial fusion. Further, temporal fusion is also introduced in BEV representation and gains great success. In this work, we propose a new method that unifies both spatial and temporal fusion and merges them into a unified mathematical formulation. The unified fusion could not only provide a new perspective on BEV fusion but also brings new capabilities. With the proposed unified spatial-temporal fusion, our method could support long-range fusion, which is hard to achieve in conventional BEV methods. Moreover, the BEV fusion in our work is temporal-adaptive and the weights of temporal fusion are learnable. In contrast, conventional methods mainly use fixed and equal weights for temporal fusion. Besides, the proposed unified fusion could avoid information lost in conventional BEV fusion methods and make full use of features. Extensive expe

riments and ablation studies on the NuScenes dataset show the effectiveness of the proposed method and our method gains the state-of-the-art performance in the map and vehicle segmentation task.
****
Point Contrastive Prediction with Semantic Clustering for Self-Supervised Learning on Point Cloud Videos
Xiaoxiao Sheng, Zhiqiang Shen, Gang Xiao, Longguang Wang, Yulan Guo, Hehe Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16515-16524
We propose a unified point cloud video self-supervised learning framework for object-centric and scene-centric data. Previous methods commonly conduct representation learning at the clip or frame level and cannot well capture fine-grained semantics. Instead of contrasting the representations of clips or frames, in this paper, we propose a unified self-supervised framework by conducting contrastive learning at the point level. Moreover, we introduce a new pretext task by achieving semantic alignment of superpoints, which further facilitates the representations to capture semantic cues at multiple scales. In addition, due to the high redundancy in the temporal dimension of dynamic point clouds, directly conducting contrastive learning at the point level usually leads to massive undesired negatives and insufficient modeling of positive representations. To remedy this, we propose a selection strategy to retain proper negatives and make use of high-similarity samples from other instances as positive supplements. Extensive experiments show that our method outperforms supervised counterparts on a wide range of downstream tasks and demonstrates the superior transferability of the learned representations.
****
Preserving Modality Structure Improves Multi-Modal Learning
Sirnam Swetha, Mamshad Nayeem Rizve, Nina Shvetsova, Hilde Kuehne, Mubarak Shah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21993-22003
Self-supervised learning on large-scale multi-modal datasets allows learning semantically meaningful embeddings in a joint multi-modal representation space without relying on human annotations. These joint embeddings enable zero-shot cross-modal tasks like retrieval and classification. However, these methods often struggle to generalize well on out-of-domain data as they ignore the semantic structure present in modality-specific embeddings. In this context, we propose a novel Semantic-Structure-Preserving Consistency approach to improve generalizability by preserving the modality-specific relationships in the joint embedding space. To capture modality-specific semantic relationships between samples, we propose to learn multiple anchors and represent the multifaceted relationship between samples with respect to their relationship with these anchors. To assign multiple anchors to each sample, we propose a novel Multi-Assignment Sinkhorn-Knopp algorithm. Our experimentation demonstrates that our proposed approach learns semantically meaningful anchors in a self-supervised manner. Furthermore, our evaluation on MSR-VTT and YouCook2 datasets demonstrates that our proposed multi-anchor assignment based solution achieves state-of-the-art performance and generalizes to both inand out-of-domain datasets. Code: https://github.com/Swetha5/Multi_Sinkhorn_Knopp
****
Focus the Discrepancy: Intra- and Inter-Correlation Learning for Image Anomaly Detection
Xincheng Yao, Ruoqi Li, Zefeng Qian, Yan Luo, Chongyang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6803-6813
Humans recognize anomalies through two aspects: larger patch-wise representation discrepancies and weaker patch-to-normal-patch correlations. However, the previous AD methods didn't sufficiently combine the two complementary aspects to design AD models. To this end, we find that Transformer can ideally satisfy the two aspects as its great power in the unified modeling of patchwise representations and patch-to-patch correlations. In this paper, we propose a novel AD framework:

FOcus-the- Discrepancy (FOD), which can simultaneously spot the patch-wise, intra- and inter-discrepancies of anomalies. The major characteristic of our method is that we renovate the self attention maps in transformers to Intra-Inter-Correlation (I2Correlation). The I2Correlation contains a two-branch structure to first explicitly establish intraand inter-image correlations, and then fuses the features of two-branch to spotlight the abnormal patterns. To learn the intra- and inter-correlations adaptively, we propose the RBF-kernel-based target-correlations as learning targets for self-supervised learning. Besides, we introduce an entropy constraint strategy to solve the mode collapse issue in optimization and further amplify the normal abnormal distinguishability. Extensive experiments on three unsupervised real-world AD benchmarks show the superior performance of our approach. Code will be available at https://github.com/xcyao00/FOD.
****
Pre-training Vision Transformers with Very Limited Synthesized Images
Ryo Nakamura, Hirokatsu Kataoka, Sora Takashima, Edgar Josafat Martinez Noriega, Rio Yokota, Nakamasa Inoue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20360-20369
Formula-driven supervised learning (FDSL) is a pre-training method that relies on synthetic images generated from mathematical formulae such as fractals.
Prior work on FDSL has shown that pre-training vision transformers on such synthetic datasets can yield competitive accuracy on a wide range of downstream tasks.
These synthetic images are categorized according to the parameters in the mathematical formula that generate them.
In the present work, we hypothesize that the process for generating different instances for the same category in FDSL, can be viewed as a form of data augmentation.
We validate this hypothesis by replacing the instances with data augmentation, which means we only need a single image per category.
Our experiments show that this one-instance fractal database (OFDB) performs better than the original dataset where instances were explicitly generated.
We further scale up OFDB to 21,000 categories and show that it matches, or even surpasses, the model pre-trained on ImageNet-21k in ImageNet-1k fine-tuning.
The number of images in OFDB is 21k, whereas ImageNet-21k has 14M.
This opens new possibilities for pre-training vision transformers with much smaller datasets.
****
Sample-adaptive Augmentation for Point Cloud Recognition Against Real-world Corruptions
Jie Wang, Lihe Ding, Tingfa Xu, Shaocong Dong, Xinli Xu, Long Bai, Jianan Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14330-14339
Robust 3D perception under corruption has become an essential task for the realm of 3D vision. While current data augmentation techniques usually perform random transformations on all point cloud objects in an offline way and ignore the structure of the samples, resulting in over-or-under enhancement. In this work, we propose an alternative to make sample-adaptive transformations based on the structure of the sample to cope with potential corruption via an auto-augmentation framework, named as AdaptPoint. Specially, we leverage a imitator, consisting of a Deformation Controller and a Mask Controller, respectively in charge of predicting deformation parameters and producing a per-point mask, based on the intrinsic structural information of the input point cloud, and then conduct corruption simulations on top. Then a discriminator is utilized to prevent the generation of excessive corruption that deviates from the original data distribution. In addition, a perception-guidance feedback mechanism is incorporated to guide the generation of samples with appropriate difficulty level. Furthermore, to address the paucity of real-world corrupted point cloud, we also introduce a new dataset ScanObjectNN-C, that exhibits greater similarity to actual data in real-world environments, especially when contrasted with preceding CAD datasets. Experiments show that our method achieves state-of-the-art results on multiple corruption ben

chmarks including ModelNet-C, our ScanObjectNN-C, and ShapeNet-C.
The source code is released at: https://github.com/Roywangj/AdaptPoint.
****
Make Encoder Great Again in 3D GAN Inversion through Geometry and Occlusion-Aware Encoding
Ziyang Yuan, Yiming Zhu, Yu Li, Hongyu Liu, Chun Yuan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2437-2447
3D GAN inversion aims to achieve high reconstruction fidelity and reasonable 3D geometry simultaneously from a single image input. However, existing 3D GAN inversion methods rely on time-consuming optimization for each individual case. In this work, we introduce a novel encoder-based inversion framework based on EG3D, one of the most widely-used 3D GAN models. We leverage the inherent properties of EG3D's latent space to design a discriminator and a background depth regularization. This enables us to train a geometry-aware encoder capable of converting the input image into corresponding latent code. Additionally, we explore the feature space of EG3D and develop an adaptive refinement stage that improves the representation ability of features in EG3D to enhance the recovery of fine-grained textural details. Finally, we propose an occlusion-aware fusion operation to prevent distortion in unobserved regions. Our method achieves impressive results comparable to optimization-based methods while operating up to 500 times faster. Our framework is well-suited for applications such as semantic editing.
****
Modality Unifying Network for Visible-Infrared Person Re-Identification
Hao Yu, Xu Cheng, Wei Peng, Weihao Liu, Guoying Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11185-11195
Visible-infrared person re-identification (VI-ReID) is a challenging task due to large cross-modality discrepancies and intra-class variations. Existing methods mainly focus on learning modality-shared representations by embedding different modalities into the same feature space. As a result, the learned feature emphasizes the common patterns across modalities while suppressing modality-specific and identity-aware information that is valuable for Re-ID. To address these issues, we propose a novel Modality Unifying Network (MUN) to explore a robust auxiliary modality for VI-ReID. First, the auxiliary modality is generated by combining the proposed cross-modality learner and intra-modality learner, which can dynamically model the modality-specific and modality-shared representations to alleviate both cross-modality and intra-modality variations. Second, by aligning identity centres across the three modalities, an identity alignment loss function is proposed to discover the discriminative feature representations. Third, a modality alignment loss is introduced to consistently reduce the distribution distance of visible and infrared images by modality prototype modeling. Extensive experiments on multiple public datasets demonstrate that the proposed method surpasses the current state-of-the-art methods by a significant margin.
****
DLT: Conditioned layout generation with Joint Discrete-Continuous Diffusion Layout Transformer
Elad Levi, Eli Brosh, Mykola Mykhailych, Meir Perez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2106-2115
Generating visual layouts is an essential ingredient of graphic design. The ability to condition layout generation on a partial subset of component attributes is critical to real-world applications that involve user interaction. Recently, diffusion models have demonstrated high-quality generative performances in various domains. However, it is unclear how to apply diffusion models to the natural representation of layouts which consists of a mix of discrete (class) and continuous (location, size) attributes. To address the conditioning layout generation problem, we introduce DLT, a joint discrete-continuous diffusion model. DLT is a transformer-based model which has a flexible conditioning mechanism that allows for conditioning on any given subset of all layout components classes, locations and sizes. Our method outperforms state-of-the-art generative models on various layout generation datasets with respect to different metrics and conditioning settings. Additionally, we validate the effectiveness of our proposed conditionin

g mechanism and the joint continuous-diffusion process. This joint process can b
e incorporated into a wide range of mixed discrete-continuous generative tasks.
****
PADDLES: Phase-Amplitude Spectrum Disentangled Early Stopping for Learning with
Noisy Labels
Huaxi Huang, Hui Kang, Sheng Liu, Olivier Salvado, Thierry Rakotoarivelo, Dadong
 Wang, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2023, pp. 16719-16730
Convolutional Neural Networks (CNNs) are powerful in learning patterns of differ
ent vision tasks, but they are sensitive to label noise and may overfit to noisy
 labels during training. The early stopping strategy averts updating CNNs during
 the early training phase and is widely employed in the presence of noisy labels
. Motivated by biological findings that the amplitude spectrum (AS) and phase sp
ectrum (PS) in the frequency domain play different roles in the animal's vision
system, we observe that PS, which captures more semantic information, can increa
se the robustness of CNNs to label noise, more so than AS can. We thus propose e
arly stops at different times for AS and PS by disentangling the features of som
e layer(s) into AS and PS using Discrete Fourier Transform (DFT) during training
. Our proposed Phase-AmplituDe DisentangLed Early Stopping (PADDLES) method is s
hown to be effective on both synthetic and real-world label-noise datasets. PADD
LES outperforms other early stopping methods and obtains state-of-the-art perfor
mance.
****
Taming Contrast Maximization for Learning Sequential, Low-latency, Event-based O
ptical Flow
Federico Paredes-Vallés, Kirk Y. W. Scheper, Christophe De Wagter, Guido C. H. E
. de Croon; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2023, pp. 9695-9705
Event cameras have recently gained significant traction since they open up new a
venues for low-latency and low-power solutions to complex computer vision proble
ms. To unlock these solutions, it is necessary to develop algorithms that can le
verage the unique nature of event data. However, the current state-of-the-art is
 still highly influenced by the frame-based literature, and usually fails to del
iver on these promises. In this work, we take this into consideration and propos
e a novel self-supervised learning pipeline for the sequential estimation of eve
nt-based optical flow that allows for the scaling of the models to high inferenc
e frequencies. At its core, we have a continuously-running stateful neural model
 that is trained using a novel formulation of contrast maximization that makes i
t robust to nonlinearities and varying statistics in the input events. Results a
cross multiple datasets confirm the effectiveness of our method, which establish
es a new state of the art in terms of accuracy for approaches trained or optimiz
ed without ground truth.
****
CLIP-Cluster: CLIP-Guided Attribute Hallucination for Face Clustering
Shuai Shen, Wanhua Li, Xiaobing Wang, Dafeng Zhang, Zhezhu Jin, Jie Zhou, Jiwen
Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC
V), 2023, pp. 20786-20795
One of the most important yet rarely studied challenges for supervised face clus
tering is the large intra-class variance caused by different face attributes suc
h as age, pose, and expression. Images of the same identity but with different f
ace attributes usually tend to be clustered into different sub-clusters. For the
 first time, we proposed an attribute hallucination framework named CLIP-Cluster
 to address this issue, which first hallucinates multiple representations for di
fferent attributes with the powerful CLIP model and then pools them by learning
neighbor-adaptive attention. Specifically, CLIP-Cluster first introduces a text-
driven attribute hallucination module, which allows one to use natural language
as the interface to hallucinate novel attributes for a given face image based on
 the well-aligned image-language CLIP space. Furthermore, we develop a neighbor-
aware proxy generator that fuses the features describing various attributes into
 a proxy feature to build a bridge among different sub-clusters and reduce the i

ntra-class variance. The proxy feature is generated by adaptively attending to t
he hallucinated visual features and the source one based on the local neighbor i
nformation. On this basis, a graph built with the proxy representations is used
for subsequent clustering operations. Extensive experiments show our proposed ap
proach outperforms state-of-the-art face clustering methods with high inference
efficiency.
****
CASSPR: Cross Attention Single Scan Place Recognition
Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, João F Henriques, Da
niel Cremers; Proceedings of the IEEE/CVF International Conference on Computer V
ision (ICCV), 2023, pp. 8461-8472
Place recognition based on point clouds (LiDAR) is an important component for au
tonomous robots or self-driving vehicles. Current SOTA performance is achieved o
n accumulated LiDAR submaps using either point-based or voxel-based structures.
While voxel-based approaches nicely integrate spatial context across multiple sc
ales, they do not exhibit the local precision of point-based methods. As a resul
t, existing methods struggle with fine-grained matching of subtle geometric feat
ures in sparse single-shot LiDAR scans. To overcome these limitations, we propos
e CASSPR as a method to fuse point-based and voxel-based approaches using cross
attention transformers. CASSPR leverages a sparse voxel branch for extracting an
d aggregating information at lower resolution and a point-wise branch for obtain
ing fine-grained local information. CASSPR uses queries from one branch to try t
o match structures in the other branch, ensuring that both extract self-containe
d descriptors of the point cloud (rather than one branch dominating), but using
both to inform the output global descriptor of the point cloud. Extensive experi
ments show that CASSPR surpasses the state-of-the-art by a large margin on sever
al datasets (Oxford RobotCar, TUM, USyd). For instance, it achieves AR@1 of 85.6
% on the TUM dataset, surpassing the strongest prior model by 15%. Our code is
publicly available.
****
DDFM: Denoising Diffusion Model for Multi-Modality Image Fusion
Zixiang Zhao, Haowen Bai, Yuanzhi Zhu, Jiangshe Zhang, Shuang Xu, Yulun Zhang, K
ai Zhang, Deyu Meng, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 8082-8093
Multi-modality image fusion aims to combine different modalities to produce fuse
d images that retain the complementary features of each modality, such as functi
onal highlights and texture details. To leverage strong generative priors and ad
dress challenges such as unstable training and lack of interpretability for GAN-
based generative methods, we propose a novel fusion algorithm based on the denoi
sing diffusion probabilistic model (DDPM). The fusion task is formulated as a co
nditional generation problem under the DDPM sampling framework, which is further
 divided into an unconditional generation subproblem and a maximum likelihood su
bproblem. The latter is modeled in a hierarchical Bayesian manner with latent va
riables and inferred by the expectation-maximization (EM) algorithm. By integrat
ing the inference solution into the diffusion sampling iteration, our method can
 generate high-quality fused images with natural image generative priors and cro
ss-modality information from source images. Note that all we required is an unco
nditional pre-trained generative model, and no fine-tuning is needed. Our extens
ive experiments indicate that our approach yields promising fusion results in in
frared-visible image fusion and medical image fusion. The code is available at h
ttps://github.com/Zhaozixiang1228/MMIF-DDFM.
****
A Unified Continual Learning Framework with General Parameter-Efficient Tuning
Qiankun Gao, Chen Zhao, Yifan Sun, Teng Xi, Gang Zhang, Bernard Ghanem, Jian Zha
ng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC
V), 2023, pp. 11483-11493
The "pre-training - downstream adaptation" presents both new opportunities and c
hallenges for Continual Learning (CL). Although the recent state-of-the-art in C
L is achieved through Parameter-Efficient-Tuning (PET) adaptation paradigm, only
 prompt has been explored, limiting its application to Transformers only. In thi

s paper, we position prompting as one instantiation of PET, and propose a unified CL framework with general PET, dubbed as Learning-Accumulation-Ensemble (LAE). PET, e.g., using Adapter, LoRA, or Prefix, can adapt a pre-trained model to downstream tasks with fewer parameters and resources. Given a PET method, our LAE framework incorporates it for CL with three novel designs. 1) Learning: the pre-trained model adapts to the new task by tuning an online PET module, along with our adaptation speed calibration to align different PET modules, 2) Accumulation: the task-specific knowledge learned by the online PET module is accumulated into an offline PET module through momentum update, 3) Ensemble: During inference, we respectively construct two experts with online/offline PET modules (which are favored by the novel/historical tasks) for prediction ensemble. We show that LAE is compatible with a battery of PET methods and gains strong CL capability. For example, LAE with Adaptor PET surpasses the prior state-of-the-art by 1.3% and 3.6% in last-incremental accuracy on CIFAR100 and ImageNet-R datasets, respectively.
****
Hierarchical Generation of Human-Object Interactions with Diffusion Probabilistic Models
Huaijin Pi, Sida Peng, Minghui Yang, Xiaowei Zhou, Hujun Bao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15061-15073
This paper presents a novel approach to generating the 3D motion of a human interacting with a target object, with a focus on solving the challenge of synthesizing long-range and diverse motions, which could not be fulfilled by existing auto-regressive models or path planning-based methods. We propose a hierarchical generation framework to solve this challenge. Specifically, our framework first generates a set of milestones and then synthesizes the motion along them. Therefore, the long-range motion generation could be reduced to synthesizing several short motion sequences guided by milestones. The experiments on the NSM, COUCH, and SAMP datasets show that our approach outperforms previous methods by a large margin in both quality and diversity. The source code is available on our project page https://zju3dv.github.io/hghoi.
****
Learning Data-Driven Vector-Quantized Degradation Model for Animation Video Super-Resolution
Zixi Tuo, Huan Yang, Jianlong Fu, Yujie Dun, Xueming Qian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13179-13189
Existing real-world video super-resolution (VSR) methods focus on designing a general degradation pipeline for open-domain videos while ignoring data intrinsic characteristics which strongly limit their performance when applying to some specific domains (e.g., animation videos). In this paper, we thoroughly explore the characteristics of animation videos and leverage the rich priors in real-world animation data for a more practical animation VSR model. In particular, we propose a multi-scale Vector-Quantized Degradation model for animation video Super-Resolution (VQD-SR) to decompose the local details from global structures and transfer the degradation priors in real-world animation videos to a learned vector-quantized codebook for degradation modeling. A rich-content Real Animation Low-quality (RAL) video dataset is collected for extracting the priors. We further propose a data enhancement strategy for high-resolution (HR) training videos based on our observation that existing HR videos are mostly collected from the Web which contains conspicuous compression artifacts. The proposed strategy is valid to lift the upper bound of animation VSR performance, regardless of the specific VSR model. Experimental results demonstrate the superiority of the proposed VQD-SR over state-of-the-art methods, through extensive quantitative and qualitative evaluations of the latest animation video super-resolution benchmark. The code and pre-trained models can be downloaded at https://github.com/researchmm/VQD-SR.
****
Compositional Feature Augmentation for Unbiased Scene Graph Generation
Lin Li, Guikun Chen, Jun Xiao, Yi Yang, Chunping Wang, Long Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2168

5-21695
Scene Graph Generation (SGG) aims to detect all the visual relation triplets <su
b, pred, obj> in a given image. With the emergence of various advanced technique
s for better utilizing both the intrinsic and extrinsic information in each rela
tion triplet, SGG has achieved great progress over the recent years. However, du
e to the ubiquitous long-tailed predicate distributions, today's SGG models are
still easily biased to the head predicates. Currently, the most prevalent debias
ing solutions for SGG are re-balancing methods, e.g., changing the distributions
 of original training samples. In this paper, we argue that all existing re-bala
ncing strategies fail to increase the diversity of the relation triplet features
 of each predicate, which is critical for robust SGG. To this end, we propose a
novel Compositional Feature Augmentation (CFA) strategy, which is the first unbi
ased SGG work to mitigate the bias issue from the perspective of increasing the
diversity of triplet features. Specifically, we first decompose each relation tr
iplet feature into two components: intrinsic feature and extrinsic feature, whic
h correspond to the intrinsic characteristics and extrinsic contexts of a relati
on triplet, respectively. Then, we design two different feature augmentation mod
ules to enrich the feature diversity of original relation triplets by replacing
or mixing up either their intrinsic or extrinsic features from other samples. Du
e to its model-agnostic nature, CFA can be seamlessly incorporated into any SGG
model. Extensive ablations have shown that CFA can achieve a new state-of-the-ar
t performance on the trade-off between different metrics.
****
Foreground and Text-lines Aware Document Image Rectification
Heng Li, Xiangping Wu, Qingcai Chen, Qianjin Xiang; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 19574-19583
This paper aims at the distorted document image rectification problem, the objec
tive to eliminate the geometric distortion in the document images and realize do
cument intelligence. Improving the readability of distorted documents is crucial
 to effectively extract information from deformed images. According to our obser
vations, the foreground and text-line of the original warped image can represent
 the deformation tendency. However, previous distorted image rectification metho
ds pay little attention to the readability of the warped paper. In this paper, w
e focus on the foreground and text-line regions of distorted paper and proposes
a global and local fusion method to improve the rectification effect of distorte
d images and enhance the readability of document images. We introduce cross atte
ntion to capture the features of the foreground and text-lines in the warped doc
ument and effectively fuse them. The proposed method is evaluated quantitatively
 and qualitatively on the public DocUNet benchmark and DIR300 Dataset, which ach
ieve state-of-the-art performances. Experimental analysis shows the proposed met
hod can well perform overall geometric rectification of distorted images and eff
ectively improve document readability (using the metrics of Character Error Rate
 and Edit Distance). The code is available at https://github.com/xiaomore/Docume
nt-Image-Dewarping.
****
Open-Vocabulary Semantic Segmentation with Decoupled One-Pass Network
Cong Han, Yujie Zhong, Dengjie Li, Kai Han, Lin Ma; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 1086-1096
Recently, the open-vocabulary semantic segmentation problem has attracted increa
sing attention and the best performing methods are based on two-stream networks:
 one stream for proposal mask generation and the other for segment classificatio
n using a pre-trained visual-language model. However, existing two-stream method
s require passing a great number of (up to a hundred) image crops into the visua
l-language model, which is highly inefficient. To address the problem, we propos
e a network that only needs a single pass through the visual-language model for
each input image. Specifically, we first propose a novel networkadaptation appro
ach, termed patch severance, to restrict the harmful interference between the pa
tch embeddings in the pre-trained visual encoder. We then propose classification
 anchor learning to encourage the network to spatially focus on more discriminat
ive features for classification. Extensive experiments demonstrate that the prop

osed method achieves outstanding performance, surpassing state-of-the-art methods while being 4 to 7 times faster at inference. Code: https://github.com/CongHan0808/DeOP.git
****

INSTA-BNN: Binary Neural Network with INSTAnce-aware Threshold

Changhun Lee, Hyungjun Kim, Eunhyeok Park, Jae-Joon Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17325-17334

Binary Neural Networks (BNNs) have emerged as a promising solution for reducing the memory footprint and compute costs of deep neural networks, but they suffer from quality degradation due to the lack of freedom as activations and weights are constrained to the binary values. To compensate for the accuracy drop, we propose a novel BNN design called Binary Neural Network with INSTAnce-aware threshold (INSTA-BNN), which controls the quantization threshold dynamically in an input-dependent or instance-aware manner. According to our observation, higher-order statistics can be a representative metric to estimate the characteristics of the input distribution. INSTA-BNN is designed to adjust the threshold dynamically considering various information, including higher-order statistics, but it is also optimized judiciously to realize minimal overhead on a real device. Our extensive study shows that INSTA-BNN outperforms the baseline by 3.0% and 2.8% on the ImageNet classification task with comparable computing cost, achieving 68.5% and 72.2% top-1 accuracy on ResNet-18 and MobileNetV1 based models, respectively.
****

Human-Inspired Facial Sketch Synthesis with Dynamic Adaptation

Fei Gao, Yifan Zhu, Chang Jiang, Nannan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7237-7247

Facial sketch synthesis (FSS) aims to generate a vivid sketch portrait from a given facial photo. Existing FSS methods merely rely on 2D representations of facial semantic or appearance. However, professional human artists usually use outlines or shadings to covey 3D geometry. Thus facial 3D geometry (e.g. depth map) is extremely important for FSS. Besides, different artists may use diverse drawing techniques and create multiple styles of sketches; but the style is globally consistent in a sketch. Inspired by such observations, in this paper, we propose a novel Human-Inspired Dynamic Adaptation (HIDA) method. Specially, we propose to dynamically modulate neuron activations based on a joint consideration of both facial 3D geometry and 2D appearance, as well as globally consistent style control. Besides, we use deformable convolutions at coarse-scales to align deep features, for generating abstract and distinct outlines. Experiments show that HIDA can generate high-quality sketches in multiple styles, and significantly outperforms previous methods, over a large range of challenging faces. Besides, HIDA allows precise style control of the synthesized sketch, and generalizes well to natural scenes and other artistic styles. Our code and results have been released online at: https://github.com/AiArt-HDU/HIDA.
****

When Epipolar Constraint Meets Non-Local Operators in Multi-View Stereo

Tianqi Liu, Xinyi Ye, Weiyue Zhao, Zhiyu Pan, Min Shi, Zhiguo Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18088-18097

Learning-based multi-view stereo (MVS) method heavily relies on feature matching, which requires distinctive and descriptive representations. An effective solution is to apply non-local feature aggregation, e.g., Transformer. Albeit useful, these techniques introduce heavy computation overheads for MVS. Each pixel densely attends to the whole image. In contrast, we propose to constrain non-local feature augmentation within a pair of lines: each point only attends the corresponding pair of epipolar lines. Our idea takes inspiration from the classic epipolar geometry, which shows that one point with different depth hypotheses will be projected to the epipolar line on the other view. This constraint reduces the 2D search space into the epipolar line in stereo matching. Similarly, this suggests that the matching of MVS is to distinguish a series of points lying on the same line. Inspired by this point-to-line search, we devise a line-to-point non-local augmentation strategy. We first devise an optimized searching algorithm to sp

lit the 2D feature maps into epipolar line pairs. Then, an Epipolar Transformer (ET) performs non-local feature augmentation among epipolar line pairs. We incorporate the ET into a learning-based MVS baseline, named ET-MVSNet. ET-MVSNet achieves state-of-the-art reconstruction performance on both the DTU and Tanks-and-Temples benchmark with high efficiency. Code is available at https://github.com/TQTQliu/ET-MVSNet.
****
LU-NeRF: Scene and Pose Estimation by Synchronizing Local Unposed NeRFs
Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, Ameesh Makadia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18312-18321
A critical obstacle preventing NeRF models from being deployed broadly in the wild is their reliance on accurate camera poses. Consequently, there is growing interest in extending NeRF models to jointly optimize camera poses and scene representation, which offers an alternative to off-the-shelf SfM pipelines which have well-understood failure modes. Existing approaches for unposed NeRF operate under limited assumptions, such as a prior pose distribution or coarse pose initialization, making them less effective in a general setting. In this work, we propose a novel approach, LU-NeRF, that jointly estimates camera poses and neural radiance fields with relaxed assumptions on pose configuration. Our approach operates in a local-to-global manner, where we first optimize over local subsets of the data, dubbed mini-scenes. LU-NeRF estimates local pose and geometry for this challenging few-shot task. The mini-scene poses are brought into a global reference frame through a robust pose synchronization step, where a final global optimization of pose and scene can be performed. We show our LU-NeRF pipeline outperforms prior attempts at unposed NeRF without making restrictive assumptions on the pose prior. This allows us to operate in the general SE(3) pose setting, unlike the baselines. Our results also indicate our model can be complementary to feature-based SfM pipelines as it compares favorably to COLMAP on low-texture and low-resolution images.
****
Calibrating Panoramic Depth Estimation for Practical Localization and Mapping
Junho Kim, Eun Sun Lee, Young Min Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8830-8840
The absolute depth values of surrounding environments provide crucial cues for various assistive technologies, such as localization, navigation, and 3D structure estimation. We propose that accurate depth estimated from panoramic images can serve as a powerful and light-weight input for a wide range of downstream tasks requiring 3D information. While panoramic images can easily capture the surrounding context from commodity devices, the estimated depth shares the limitations of conventional image-based depth estimation; the performance deteriorates under large domain shifts and the absolute values are still ambiguous to infer from 2D observations. By taking advantage of the holistic view, we mitigate such effects in a self-supervised way and fine-tune the network with geometric consistency during the test phase. Specifically, we construct a 3D point cloud from the current depth prediction and project the point cloud at various viewpoints or apply stretches on the current input image to generate synthetic panoramas. Then we minimize the discrepancy of the 3D structure estimated from synthetic images without collecting additional data. We empirically evaluate our method in robot navigation and map-free localization where our method shows large performance enhancements. Our calibration method can therefore widen the applicability under various external conditions, serving as a key component for practical panorama-based machine vision systems.
****
DiffDis: Empowering Generative Diffusion Model with Cross-Modal Discrimination Capability
Runhui Huang, Jianhua Han, Guansong Lu, Xiaodan Liang, Yihan Zeng, Wei Zhang, Hang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15713-15723
Recently, large-scale diffusion models, e.g., Stable diffusion and DallE2, have

shown remarkable results on image synthesis. On the other hand, large-scale cross-modal pre-trained models (e.g., CLIP, ALIGN, and FILIP) are competent for various downstream tasks by learning to align vision and language embeddings. In this paper, we explore the possibility of jointly modeling generation and discrimination. Specifically, we propose DiffDis to unify the cross-modal generative and discriminative pretraining into one single framework under the diffusion process. DiffDis first formulates the image-text discriminative problem as a generative diffusion process of the text embedding from the text encoder conditioned on the image. Then, we propose a novel dual-stream network architecture, which fuses the noisy text embedding with the knowledge of latent images from different scales for image-text discriminative learning. Moreover, the generative and discriminative tasks can efficiently share the image-branch network structure in the multi-modality model. Benefiting from diffusion-based unified training, DiffDis achieves both better generation ability and cross-modal semantic alignment in one architecture. Experimental results show that DiffDis outperforms single-task models on both the image generation and the image-text discriminative tasks, e.g., 1.65% improvement on average accuracy of zero-shot classification over 12 datasets and 2.42 improvement on FID of zero-shot image synthesis.
****
DS-Fusion: Artistic Typography via Discriminated and Stylized Diffusion
Maham Tanveer, Yizhi Wang, Ali Mahdavi-Amiri, Hao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 374-384
We introduce a novel method to automatically generate an artistic typography by stylizing one or more letter fonts to visually convey the semantics of an input word, while ensuring that the output remains readable. To address an assortment of challenges with our task at hand including conflicting goals (artistic stylization vs. legibility), lack of ground truth, and immense search space, our approach utilizes large language models to bridge texts and visual images for stylization and build an unsupervised generative model with a diffusion model backbone. Specifically, we employ the denoising generator in Latent Diffusion Model (LDM), with the key addition of a CNN-based discriminator to adapt the input style onto the input text. The discriminator uses rasterized images of a given letter/word font as real samples and the output of the denoising generator as fake samples. Our model is coined DS-Fusion for discriminated and stylized diffusion. We showcase the quality and versatility of our method through numerous examples, qualitative and quantitative evaluation, and ablation studies. User studies comparing to strong baselines including CLIPDraw, DALL-E 2, Stable Diffusion, as well as artist-crafted typographies, demonstrate strong performance of DS-Fusion.
****
Distilling DETR with Visual-Linguistic Knowledge for Open-Vocabulary Object Detection
Liangqi Li, Jiaxu Miao, Dahu Shi, Wenming Tan, Ye Ren, Yi Yang, Shiliang Pu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6501-6510
Current methods for open-vocabulary object detection (OVOD) rely on a pre-trained vision-language model (VLM) to acquire the recognition ability. In this paper, we propose a simple yet effective framework to Distill the Knowledge from the VLM to a DETR-like detector, termed DK-DETR. Specifically, we present two ingenious distillation schemes named semantic knowledge distillation (SKD) and relational knowledge distillation (RKD). To utilize the rich knowledge from the VLM systematically, SKD transfers the semantic knowledge explicitly, while RKD exploits implicit relationship information between objects. Furthermore, a distillation branch including a group of auxiliary queries is added to the detector to mitigate the negative effect on base categories. Equipped with SKD and RKD on the distillation branch, DK-DETR improves the detection performance of novel categories significantly and avoids disturbing the detection of base categories. Extensive experiments on LVIS and COCO datasets show that DK-DETR surpasses existing OVOD methods under the setting that the base-category supervision is solely available. The code and models are available at https://github.com/hikvision-research/opera.

****
Scale-MAE: A Scale-Aware Masked Autoencoder for Multiscale Geospatial Representa
tion Learning
Colorado J Reed, Ritwik Gupta, Shufan Li, Sarah Brockman, Christopher Funk, Bria
n Clipp, Kurt Keutzer, Salvatore Candido, Matt Uyttendaele, Trevor Darrell; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023
, pp. 4088-4099
Large, pretrained models are commonly finetuned with imagery that is heavily aug
mented to mimic different conditions and scales, with the resulting models used
for various tasks with imagery from a range of spatial scales. Such models overl
ook scale-specific information in the data for scale-dependent domains, such as
remote sensing. In this paper, we present Scale-MAE, a pretraining method that e
xplicitly learns relationships between data at different, known scales throughou
t the pretraining process. Scale-MAE pretrains a network by masking an input ima
ge at a known input scale, where the area of the Earth covered by the image dete
rmines the scale of the ViT positional encoding, not the image resolution. Scale
-MAE encodes the masked image with a standard ViT backbone, and then decodes the
 masked image through a bandpass filter to reconstruct low/high frequency images
 at lower/higher scales. We find that tasking the network with reconstructing bo
th low/high frequency images leads to robust multiscale representations for remo
te sensing imagery. Scale-MAE achieves an average of a 2.4 - 5.6% non-parametric
 kNN classification improvement across eight remote sensing datasets compared to
 current state-of-the-art and obtains a 0.9 mIoU to 1.7 mIoU improvement on the
SpaceNet building segmentation transfer task for a range of evaluation scales.
****
View Consistent Purification for Accurate Cross-View Localization
Shan Wang, Yanhao Zhang, Akhil Perincherry, Ankit Vora, Hongdong Li; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8
197-8206
This paper proposes a fine-grained self-localization method for outdoor robotics
 that utilizes a flexible number of onboard cameras and readily accessible satel
lite images. The proposed method addresses limitations in existing cross-view lo
calization methods that struggle to handle noise sources such as moving objects
and seasonal variations. It is the first sparse visual-only method that enhances
 perception in dynamic environments by detecting view-consistent key points and
their corresponding deep features from ground and satellite views, while removin
g off-the-ground objects and establishing homography transformation between the
two views.
 Moreover, the proposed method incorporates a spatial embedding approach that le
verages camera intrinsic and extrinsic information to reduce the ambiguity of pu
rely visual matching, leading to improved feature matching and overall pose esti
mation accuracy. The method exhibits strong generalization and is robust to envi
ronmental changes, requiring only geo-poses as ground truth. Extensive experimen
ts on the KITTI and Ford Multi-AV Seasonal datasets demonstrate that our propose
d method outperforms existing state-of the-art methods, achieving median spatial
 accuracy errors below 0.5 meters along the lateral and longitudinal directions,
 and a median orientation accuracy error below 2 degrees.
****
A Unified Framework for Robustness on Diverse Sampling Errors
Myeongho Jeon, Myungjoo Kang, Joonseok Lee; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 1464-1472
Recent studies have substantiated that machine learning algorithms including con
volutional neural networks often suffer from unreliable generalizations when the
re is a significant gap between the source and target data distributions. To mit
igate this issue, a predetermined distribution shift has been addressed independ
ently (e.g., single domain generalization, de-biasing). However, a distribution
mismatch cannot be clearly estimated because the target distribution is unknown
at training. Therefore, a conservative approach robust on unexpected diverse dis
tributions is more desirable in practice. Our work starts from a motivation to a
llow adaptive inference once we know the target, since it is accessible only at

testing. Instead of assuming and fixing the target distribution at training, our
 proposed approach allows adjusting the feature space the model refers to at eve
ry prediction, i.e., instance-wise adaptive inference. The extensive evaluation
demonstrates our method is effective for generalization on diverse distributions
.
****
Efficient Video Action Detection with Token Dropout and Context Refinement
Lei Chen, Zhan Tong, Yibing Song, Gangshan Wu, Limin Wang; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10388-10399
Streaming video clips with large-scale video tokens impede vision transformers (
ViTs) for efficient recognition, especially in video action detection where suff
icient spatiotemporal representations are required for precise actor identificat
ion. In this work, we propose an end-to-end framework for efficient video action
 detection (EVAD) based on vanilla ViTs. Our EVAD consists of two specialized de
signs for video action detection. First, we propose a spatiotemporal token dropo
ut from a keyframe-centric perspective. In a video clip, we maintain all tokens
from its keyframe, preserve tokens relevant to actor motions from other frames,
and drop out the remaining tokens in this clip. Second, we refine scene context
by leveraging remaining tokens for better recognizing actor identities. The regi
on of interest (RoI) in our action detector is expanded into temporal domain. Th
e captured spatiotemporal actor identity representations are refined via scene c
ontext in a decoder with the attention mechanism. These two designs make our EVA
D efficient while maintaining accuracy, which is validated on three benchmark da
tasets (i.e., AVA, UCF101-24, JHMDB). Compared to the vanilla ViT backbone, our
EVAD reduces the overall GFLOPs by 43% and improves real-time inference speed by
 40% with no performance degradation. Moreover, even at similar computational co
sts, our EVAD can improve the performance by 1.1 mAP with higher resolution inpu
ts. Code is available at https://github.com/MCG-NJU/EVAD.
****
Explicit Motion Disentangling for Efficient Optical Flow Estimation
Changxing Deng, Ao Luo, Haibin Huang, Shaodan Ma, Jiangyu Liu, Shuaicheng Liu; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 9521-9530
In this paper, we propose a novel framework for optical flow estimation that ach
ieves a good balance between performance and efficiency. Our approach involves d
isentangling global motion learning from local flow estimation, treating global
matching and local refinement as separate stages. We offer two key insights: Fir
st, the multi-scale 4D cost-volume based recurrent flow decoder is computational
ly expensive and unnecessary for handling small displacement. With the separatio
n, we can utilize lightweight methods for both parts and maintain similar perfor
mance. Second, a dense and robust global matching is essential for both flow ini
tialization as well as stable and fast convergence for the refinement stage. Tow
ards this end, we introduce EMD-Flow, a framework that explicitly separates glob
al motion estimation from the recurrent refinement stage. We propose two novel m
odules: Multi-scale Motion Aggregation (MMA) and Confidence-induced Flow Propaga
tion (CFP). These modules leverage cross-scale matching prior and self-contained
 confidence maps to handle the ambiguities of dense matching in a global manner,
 generating a dense initial flow. Additionally, a lightweight decoding module is
 followed to handle small displacements, resulting in an efficient yet robust fl
ow estimation framework. We further conduct comprehensive experiments on standar
d optical flow benchmarks with the proposed framework, and the experimental resu
lts demonstrate its superior balance between performance and runtime. Code is av
ailable at https://github.com/gddcx/EMD-Flow.
****
LiDAR-Camera Panoptic Segmentation via Geometry-Consistent and Semantic-Aware Al
ignment
Zhiwei Zhang, Zhizhong Zhang, Qian Yu, Ran Yi, Yuan Xie, Lizhuang Ma; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
3662-3671
3D panoptic segmentation is a challenging perception task that requires both sem

antic segmentation and instance segmentation. In this task, we notice that image
s could provide rich texture, color, and discriminative information, which can c
omplement LiDAR data for evident performance improvement, but their fusion remai
ns a challenging problem. To this end, we propose LCPS, the first LiDAR-Camera P
anoptic Segmentation network. In our approach, we conduct LiDAR-Camera fusion in
 three stages: 1) an Asynchronous Compensation Pixel Alignment (ACPA) module tha
t calibrates the coordinate misalignment caused by asynchronous problems between
 sensors; 2) a Semantic-Aware Region Alignment (SARA) module that extends the on
e-to-one point-pixel mapping to one-to-many semantic relations; 3) a Point-to-Vo
xel feature Propagation (PVP) module that integrates both geometric and semantic
 fusion information for the entire point cloud. Our fusion strategy improves abo
ut 6.9% PQ performance over the LiDAR-only baseline on NuScenes dataset. Extensi
ve quantitative and qualitative experiments further demonstrate the effectivenes
s of our novel framework. The code will be released at https://github.com/zhangz
w12319/lcps.git.
****
GrowCLIP: Data-Aware Automatic Model Growing for Large-scale Contrastive Languag
e-Image Pre-Training
Xinchi Deng, Han Shi, Runhui Huang, Changlin Li, Hang Xu, Jianhua Han, James Kwo
k, Shen Zhao, Wei Zhang, Xiaodan Liang; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2023, pp. 22178-22189
Cross-modal pre-training has shown impressive performance on a wide range of dow
nstream tasks, benefiting from massive image-text pairs collected from the Inter
net. In practice, online data are growing constantly, highlighting the importanc
e of the ability of pre-trained model to learn from data that is continuously gr
owing. Existing works on cross-modal pre-training mainly focus on training a net
work with fixed architecture. However, it is impractical to limit the model capa
city when considering the continuously growing nature of pre-training data in re
al-world applications. On the other hand, it is important to utilize the knowled
ge in current model to obtain efficient training and better performance. To addr
ess the above issues, in this paper, we propose GrowCLIP, a data-driven automati
c model growing algorithm for contrastive language-image pre-training with conti
nuous image-text pairs as input. Specially, we adopt a dynamic growth space and
seek out the optimal architecture at each growth step to adapt to online learnin
g scenarios. And the shared encoder is proposed in our growth space to enhance t
he degree of cross-modal fusion. Besides, we explore the effect of growth in dif
ferent dimensions, which could provide future references for the design of cross
-modal model architecture. Finally, we employ parameter inheriting with momentum
 (PIM) to maintain the previous knowledge and address the issue of local minimum
 dilemma. Compared with the existing methods, GrowCLIP improve 2.3% average top-
1 accuracy on zero-shot image classification of 9 downstream tasks. As for zero-
shot image retrieval, GrowCLIP can improve 1.2% for top-1 image-to-text recall o
n Flickr30K dataset.
****
From Chaos Comes Order: Ordering Event Representations for Object Recognition an
d Detection
Nikola Zubi■, Daniel Gehrig, Mathias Gehrig, Davide Scaramuzza; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12846-
12856
Today, state-of-the-art deep neural networks that process events first convert t
hem into dense, grid-like input representations before using an off-the-shelf ne
twork. However, selecting the appropriate representation for the task traditiona
lly requires training a neural network for each representation and selecting the
 best one based on the validation score, which is very time-consuming. This work
 eliminates this bottleneck by selecting representations based on the Gromov-Was
serstein Discrepancy (GWD) between raw events and their representation. It is ab
out 200 times faster to compute than training a neural network and preserves the
 task performance ranking of event representations across multiple representatio
ns, network backbones, datasets, and tasks. Thus finding representations with hi
gh task scores is equivalent to finding representations with a low GWD. We use t

his insight to, for the first time, perform a hyperparameter search on a large family of event representations, revealing new and powerful representations that exceed the state-of-the-art. Our optimized representations outperform existing representations by 1.7 mAP on the 1 Mpx dataset and 0.3 mAP on the Gen1 dataset, two established object detection benchmarks, and reach a 3.8% higher classification score on the mini N-ImageNet benchmark. Moreover, we outperform state-of-the-art by 2.1 mAP on Gen1 and state-of-the-art feed-forward methods by 6.0 mAP on the 1 Mpx datasets. This work opens a new unexplored field of explicit representation optimization for event-based learning.
****
LA-Net: Landmark-Aware Learning for Reliable Facial Expression Recognition under Label Noise
Zhiyu Wu, Jinshi Cui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20698-20707
Facial expression recognition (FER) remains a challenging task due to the ambiguity of expressions. The derived noisy labels significantly harm the performance in real-world scenarios. To address this issue, we present a new FER model named Landmark-Aware Net (LA-Net), which leverages facial landmarks to mitigate the impact of label noise from two perspectives. Firstly, LA-Net uses landmark information to suppress the uncertainty in expression space and constructs the label distribution of each sample by neighborhood aggregation, which in turn improves the quality of training supervision. Secondly, the model incorporates landmark information into expression representations using the devised expression-landmark contrastive loss. The enhanced expression feature extractor can be less susceptible to label noise. Our method can be integrated with any deep neural network for better training supervision without introducing extra inference costs. We conduct extensive experiments on both in-the-wild datasets and synthetic noisy datasets and demonstrate that LA-Net achieves state-of-the-art performance.
****
Identity-Consistent Aggregation for Video Object Detection
Chaorui Deng, Da Chen, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13434-13444
In Video Object Detection (VID), a common practice is to leverage the rich temporal contexts from the video to enhance the object representations in each frame. Existing methods treat the temporal contexts obtained from different objects indiscriminately and ignore their different identities. While intuitively, aggregating local views of the same object in different frames may facilitate a better understanding of the object. Thus, in this paper, we aim to enable the model to focus on the identity-consistent temporal contexts of each object to obtain more comprehensive object representations and handle the rapid object appearance variations such as occlusion, motion blur, etc. However, realizing this goal on top of existing VID models faces low-efficiency problems due to their redundant region proposals and nonparallel frame-wise prediction manner. To aid this, we propose ClipVID, a VID model equipped with Identity-Consistent Aggregation (ICA) layers specifically designed for mining fine-grained and identity-consistent temporal contexts. It effectively reduces the redundancies through the set prediction strategy, making the ICA layers very efficient and further allowing us to design an architecture that makes parallel clip-wise predictions for the whole video clip. Extensive experimental results demonstrate the superiority of our method: a state-of-the-art (SOTA) performance (84.7% mAP) on the ImageNet VID dataset while running at a speed about 7x faster (39.3 fps) than previous SOTAs.
****
Scene-Aware Label Graph Learning for Multi-Label Image Classification
Xuelin Zhu, Jian Liu, Weijia Liu, Jiawei Ge, Bo Liu, Jiuxin Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1473-1482
Multi-label image classification refers to assigning a set of labels for an image. One of the main challenges of this task is how to effectively capture the correlation among labels. Existing studies on this issue mostly rely on the statistical label co-occurrence or semantic similarity of labels. However, an important

fact is ignored that the co-occurrence of labels is closely related with image scenes (indoor, outdoor, etc.), which is a vital characteristic in multi-label image classification. In this paper, a novel scene-aware label graph learning framework is proposed, which is capable of learning visual representations for labels while fully perceiving their co-occurrence relationships under variable scenes. Specifically, our framework is able to detect scene categories of images without relying on manual annotations, and keeps track of the co-occurring labels by maintaining a global co-occurrence matrix for each scene category throughout the whole training phase. These scene-independent co-occurrence matrices are further employed to guide the interactions among label representations in a graph propagation manner towards accurate label prediction. Extensive experiments on public benchmarks demonstrate the superiority of our proposed framework compared to the state of the arts. Code will be publicly available soon.
****
Relightify: Relightable 3D Faces from a Single Image via Diffusion Models
Foivos Paraperas Papantoniou, Alexandros Lattas, Stylianos Moschoglou, Stefanos Zafeiriou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8806-8817
Following the remarkable success of diffusion models on image generation, recent works have also demonstrated their impressive ability to address a number of inverse problems in an unsupervised way, by properly constraining the sampling process based on a conditioning input. Motivated by this, in this paper, we present the first approach to use diffusion models as a prior for highly accurate 3D facial BRDF reconstruction from a single image. We start by leveraging a high-quality UV dataset of facial reflectance (diffuse and specular albedo and normals), which we render under varying illumination settings to simulate natural RGB textures and, then, train an unconditional diffusion model on concatenated pairs of rendered textures and reflectance components. At test time, we fit a 3D morphable model to the given image and unwrap the face in a partial UV texture. By sampling from the diffusion model, while retaining the observed texture part intact, the model inpaints not only the self-occluded areas but also the unknown reflectance components, in a single sequence of denoising steps. In contrast to existing methods, we directly acquire the observed texture from the input image, thus, resulting in more faithful and consistent reflectance estimation. Through a series of qualitative and quantitative comparisons, we demonstrate superior performance in both texture completion as well as reflectance reconstruction tasks.
****
Fcaformer: Forward Cross Attention in Hybrid Vision Transformer
Haokui Zhang, Wenze Hu, Xiaoyu Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6060-6069
Currently, one main research line in designing more efficient vision transformer is reducing computational cost of self attention modules by adopting sparse attention or using local attention windows. In contrast, we propose a different approach that aims to improve the performance of transformer-based architectures by densifying the attention pattern. Specifically, we proposed forward cross attention for hybrid vision transformer (FcaFormer), where tokens from previous blocks in the same stage are secondary used. To achieve this, the FcaFormer leverages two innovative components: learnable scale factors (LSFs) and a token merge and enhancement module (TME). The LSFs enable efficient processing of cross tokens, while the TME generates representative cross tokens. By integrating these components, the proposed FcaFormer enhances the interactions of tokens across blocks with potentially different semantics, and encourages more information flows to the lower levels. Based on the forward cross attention (Fca), we have designed a series of FcaFormer models that achieve the best trade-off between model size, computational cost, memory cost, and accuracy. For example, without the need for knowledge distillation to strengthen training, our FcaFormer achieves 83.1% top-1 accuracy on Imagenet with only 16.3 million parameters and about 3.6 billion MACs. This saves almost half of the parameters and a few computational cost while achieving 0.7% higher accuracy compared with distilled EfficientFormer
****

Progressive Spatio-Temporal Prototype Matching for Text-Video Retrieval
Pandeng Li, Chen-Wei Xie, Liming Zhao, Hongtao Xie, Jiannan Ge, Yun Zheng, Deli Zhao, Yongdong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4100-4110
The performance of text-video retrieval has been significantly improved by vision-language cross-modal learning schemes.
The typical solution is to directly align the global video-level and sentence-level features during learning, which would ignore the intrinsic video-text relations, i.e., a text description only corresponds to a spatio-temporal part of videos.
Hence, the matching process should consider both fine-grained spatial content and various temporal semantic events.
To this end, we propose a text-video learning framework with progressive spatio-temporal prototype matching. Specifically, the vanilla matching process is decomposed into two complementary phases: object-phrase prototype matching and event-sentence prototype matching. In the object-phrase prototype matching phase, a spatial prototype generation mechanism is developed to predict key patches or words, which are sparsely integrated into object or phrase prototypes. Importantly, optimizing the local alignment between object-phrase prototypes helps the model perceive spatial details. In the event-sentence prototype matching phase, we design a temporal prototype generation mechanism to associate intra-frame objects and interact inter-frame temporal relations. Such progressively generated event prototypes can reveal semantic diversity in videos for dynamic matching. Validated by comprehensive experiments, our method consistently outperforms the state-of-the-art methods on four video retrieval benchmarks.
****
Leveraging Spatio-Temporal Dependency for Skeleton-Based Action Recognition
Jungho Lee, Minhyeok Lee, Suhwan Cho, Sungmin Woo, Sungjun Jang, Sangyoun Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10255-10264
Skeleton-based action recognition has attracted considerable attention due to its compact representation of the human body's skeletal sructure. Many recent methods have achieved remarkable performance using graph convolutional networks (GCNs) and convolutional neural networks (CNNs), which extract spatial and temporal features, respectively. Although spatial and temporal dependencies in the human skeleton have been explored separately, spatio-temporal dependency is rarely considered. In this paper, we propose the Spatio-Temporal Curve Network (STC-Net) to effectively leverage the spatio-temporal dependency of the human skeleton. Our proposed network consists of two novel elements: 1) The Spatio-Temporal Curve (STC) module; and 2) Dilated Kernels for Graph Convolution (DK-GC). The STC module dynamically adjusts the receptive field by identifying meaningful node connections between every adjacent frame and generating spatio-temporal curves based on the identified node connections, providing an adaptive spatio-temporal coverage. In addition, we propose DK-GC to consider long-range dependencies, which results in a large receptive field without any additional parameters by applying an extended kernel to the given adjacency matrices of the graph. Our STC-Net combines these two modules and achieves state-of-the-art performance on four skeleton-based action recognition benchmarks.
****
Data Augmented Flatness-aware Gradient Projection for Continual Learning
Enneng Yang, Li Shen, Zhenyi Wang, Shiwei Liu, Guibing Guo, Xingwei Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5630-5639
The goal of continual learning (CL) is to continuously learn new tasks without forgetting previously learned old tasks. To alleviate catastrophic forgetting, gradient projection based CL methods require that the gradient updates of new tasks are orthogonal to the subspace spanned by old tasks. This limits the learning process and leads to poor performance on the new task due to the projection constraint being too strong. In this paper, we first revisit the gradient projection method from the perspective of flatness of loss surface, and find that unflatne

ss of the loss surface leads to catastrophic forgetting of the old tasks when th
e projection constraint is reduced to improve the performance of new tasks. Base
d on our findings, we propose a Data Augmented Flatness-aware Gradient Projectio
n (DFGP) method to solve the problem, which consists of three modules: data and
weight perturbation, flatness-aware optimization, and gradient projection. Speci
fically, we first perform a flatness-aware perturbation on the task data and cur
rent weights to find the case that makes the task loss worst. Next, flatness-awa
re optimization optimizes both the loss and the flatness of the loss surface on
raw and worst-case perturbed data to obtain a flatness-aware gradient. Finally,
gradient projection updates the network with the flatness-aware gradient along d
irections orthogonal to the subspace of the old tasks. Extensive experiments on
four datasets show that our method improves the flatness of loss surface and the
 performance of new tasks, and achieves state-of-the-art (SOTA) performance in t
he average accuracy of all tasks.
****
Camera-Driven Representation Learning for Unsupervised Domain Adaptive Person Re
-identification
Geon Lee, Sanghoon Lee, Dohyung Kim, Younghoon Shin, Yongsang Yoon, Bumsub Ham;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2023, pp. 11453-11462
We present a novel unsupervised domain adaption method for person re-identificat
ion (reID) that generalizes a model trained on a labeled source domain to an unl
abeled target domain. We introduce a camera-driven curriculum learning (CaCL) fr
amework that leverages camera labels of person images to transfer knowledge from
 source to target domains progressively. To this end, we divide target domain da
taset into multiple subsets based on the camera labels, and initially train our
model with a single subset (i.e., images captured by a single camera). We then g
radually exploit more subsets for training, according to a curriculum sequence o
btained with a camera-driven scheduling rule. The scheduler considers maximum me
an discrepancies (MMD) between each subset and the source domain dataset, such t
hat the subset closer to the source domain is exploited earlier within the curri
culum. For each curriculum sequence, we generate pseudo labels of person images
in a target domain to train a reID model in a supervised way. We have observed t
hat the pseudo labels are highly biased toward cameras, suggesting that person i
mages obtained from the same camera are likely to have the same pseudo labels, e
ven for different IDs. To address the camera bias problem, we also introduce a c
amera-diversity (CD) loss encouraging person images of the same pseudo label, bu
t captured across various cameras, to involve more for discriminative feature le
arning, providing person representations robust to inter-camera variations. Expe
rimental results on standard benchmarks, including real-to-real and synthetic-to
-real scenarios, demonstrate the effectiveness of our framework.
****
Sample-wise Label Confidence Incorporation for Learning with Noisy Labels
Chanho Ahn, Kikyung Kim, Ji-won Baek, Jongin Lim, Seungju Han; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1823-18
32
Deep learning algorithms require large amounts of labeled data for effective per
formance, but the presence of noisy labels often significantly degrade their per
formance. Although recent studies on designing a robust objective function to la
bel noise, known as the robust loss method, have shown promising results for lea
rning with noisy labels, they suffer from the issue of underfitting not only noi
sy samples but also clean ones, leading to suboptimal model performance. To addr
ess this issue, we propose a novel learning framework that selectively suppresse
s noisy samples while avoiding underfitting clean data. Our framework incorporat
es label confidence as a measure of label noise, enabling the network model to p
rioritize the training of samples deemed to be noise-free. The label confidence
is based on the robust loss methods, and we provide theoretical evidence that ou
r method can reach the optimal point of the robust loss, subject to certain cond
itions. Furthermore, the proposed method is generalizable and can be combined wi
th existing robust loss methods, making it suitable for a wide range of applicat

ions of learning with noisy labels. We evaluate our approach on both synthetic and real-world datasets, and the experimental results demonstrate its effectiveness in achieving outstanding classification performance compared to state-of-the-art methods.
****
CLIPTrans: Transferring Visual Knowledge with Pre-trained Models for Multimodal Machine Translation
Devaansh Gupta, Siddhant Kharbanda, Jiawei Zhou, Wanhua Li, Hanspeter Pfister, Donglai Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2875-2886
There has been a growing interest in developing multimodal machine translation (MMT) systems that enhance neural machine translation (NMT) with visual knowledge. This problem setup involves using images as auxiliary information during training, and more recently, eliminating their use during inference. Towards this end, previous works face a challenge in training powerful MMT models from scratch due to the scarcity of annotated multilingual vision-language data, especially for low-resource languages. Simultaneously, there has been an influx of multilingual pre-trained models for NMT and multimodal pre-trained models for vision-language tasks, primarily in English, which have shown exceptional generalisation ability. However, these are not directly applicable to MMT since they do not provide aligned multimodal multilingual features for generative tasks. To alleviate this issue, instead of designing complex modules for MMT, we propose CLIPTrans, which simply adapts the independently pre-trained multimodal M-CLIP and the multilingual mBART. In order to align their embedding spaces, mBART is conditioned on the M-CLIP features by a prefix sequence generated through a lightweight mapping network. We train this in a two-stage pipeline which warms up the model with image captioning before the actual translation task. Through experiments, we demonstrate the merits of this framework and consequently push forward the state-of-the-art across standard benchmarks by an average of +2.67 BLEU. The code can be found at www.github.com/devaansh100/CLIPTrans.
****
SGAligner: 3D Scene Alignment with Scene Graphs
Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, Iro Armeni; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21927-21937
Building 3D scene graphs has recently emerged as a topic in scene representation for several embodied AI applications to represent the world in a structured and rich manner. With their increased use in solving downstream tasks (e.g., navigation and room rearrangement), can we leverage and recycle them for creating 3D maps of environments, a pivotal step in agent operation? We focus on the fundamental problem of aligning pairs of 3D scene graphs whose overlap can range from zero to partial and can contain arbitrary changes. We propose SGAligner, the first method for aligning pairs of 3D scene graphs that is robust to in-the-wild scenarios (i.e., unknown overlap - if any - and changes in the environment). We get inspired by multi-modality knowledge graphs and use contrastive learning to learn a joint, multi-modal embedding space. We evaluate on the 3RScan dataset and further showcase that our method can be used for estimating the transformation between pairs of 3D scenes. Since benchmarks for these tasks are missing, we create them on this dataset. The code, benchmark, and trained models are available on the project website.
****
Name Your Colour For the Task: Artificially Discover Colour Naming via Colour Quantisation Transformer
Shenghan Su, Lin Gu, Yue Yang, Zenghui Zhang, Tatsuya Harada; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12021-12031
The long-standing theory that a colour-naming system evolves under dual pressure of efficient communication and perceptual mechanism is supported by more and more linguistic studies, including analysing four decades of diachronic data from the Nafaanra language. This inspires us to explore whether machine learning coul

d evolve and discover a similar colour-naming system via optimising the communic ation efficiency represented by high-level recognition performance. Here, we pro pose a novel colour quantisation transformer, CQFormer, that quantises colour sp ace while maintaining the accuracy of machine recognition on the quantised image s. Given an RGB image, Annotation Branch maps it into an index map before genera ting the quantised image with a colour palette; meanwhile the Palette Branch uti lises a key-point detection way to find proper colours in the palette among the whole colour space. By interacting with colour annotation, CQFormer is able to b alance both the machine vision accuracy and colour perceptual structure such as distinct and stable colour distribution for discovered colour system. Very inter estingly, we even observe the consistent evolution pattern between our artificia l colour system and basic colour terms across human languages. Besides, our colo ur quantisation method also offers an efficient quantisation method that effecti vely compresses the image storage while maintaining high performance in high-lev el recognition tasks such as classification and detection. Extensive experiments demonstrate the superior performance of our method with extremely low bit-rate colours, showing potential to integrate into quantisation network to quantities from image to network activation. The source code is available at https://github .com/ryeocthiv/CQFormer.
****
FSAR: Federated Skeleton-based Action Recognition with Adaptive Topology Structu re and Knowledge Distillation
Jingwen Guo, Hong Liu, Shitong Sun, Tianyu Guo, Min Zhang, Chenyang Si; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp . 10400-10410
Existing skeleton-based action recognition methods typically follow a centralize d learning paradigm, which can pose privacy concerns when exposing human-related videos. Federated Learning (FL) has attracted much attention due to its outstan ding advantages in privacy-preserving. However, directly applying FL approaches to skeleton videos suffers from unstable training. In this paper, we investigate and discover that the heterogeneous human topology graph structure is the cruci al factor hindering training stability. To address this issue, we pioneer a nove l Federated Skeleton-based Action Recognition (FSAR) paradigm, which enables the construction of a globally generalized model without accessing local sensitive data. Specifically, we introduce an Adaptive Topology Structure (ATS), separatin g generalization and personalization by learning a domain-invariant topology sha red across clients and a domain-specific topology decoupled from global model ag gregation. Furthermore, we explore Multi-grain Knowledge Distillation (MKD) to m itigate the discrepancy between clients and the server caused by distinct updati ng patterns through aligning shallow block-wise motion features. Extensive exper iments on multiple datasets demonstrate that FSAR outperforms state-of-the-art F L-based methods while inherently protecting privacy for skeleton-based action re cognition.
****
Video Adverse-Weather-Component Suppression Network via Weather Messenger and Ad versarial Backpropagation
Yijun Yang, Angelica I. Aviles-Rivero, Huazhu Fu, Ye Liu, Weiming Wang, Lei Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13200-13210
Although convolutional neural networks (CNNs) have been proposed to remove adver se weather conditions in single images using a single set of pre-trained weights , they fail to restore weather videos due to the absence of temporal information . Furthermore, existing methods for removing adverse weather conditions (e.g., r ain, fog, and snow) from videos can only handle one type of adverse weather. In this work, we propose the first framework for restoring videos from all adverse weather conditions by developing a video adverse-weather-component suppression n etwork (ViWS-Net). To achieve this, we first devise a weather-agnostic video tra nsformer encoder with multiple transformer stages. Moreover, we design a long sh ort-term temporal modeling mechanism for weather messenger to early fuse input a djacent video frames and learn weather-specific information. We further introduc

e a weather discriminator with gradient reversion, to maintain the weather-invariant common information and suppress the weather-specific information in pixel features, by adversarially predicting weather types. Finally, we develop a messenger-driven video transformer decoder to retrieve the residual weather-specific feature, which is spatiotemporally aggregated with hierarchical pixel features and refined to predict the clean target frame of input videos. Experimental results, on benchmark datasets and real-world weather videos, demonstrate that our ViWS-Net outperforms current state-of-the-art methods in terms of restoring videos degraded by any weather condition.
****
Efficient Discovery and Effective Evaluation of Visual Perceptual Similarity: A Benchmark and Beyond
Oren Barkan, Tal Reiss, Jonathan Weill, Ori Katz, Roy Hirsch, Itzik Malkiel, Noam Koenigstein; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20007-20018
Visual similarities discovery (VSD) is an important task with broad e-commerce applications. Given an image of a certain object, the goal of VSD is to retrieve images of different objects with high perceptual visual similarity. Although being a highly addressed problem, the evaluation of proposed methods for VSD is often based on a proxy of an identification-retrieval task, evaluating the ability of a model to retrieve different images of the same object. We posit that evaluating VSD methods based on identification tasks is limited, and faithful evaluation must rely on expert annotations. In this paper, we introduce the first large-scale fashion visual similarity benchmark dataset, consisting of more than 110K expert-annotated image pairs. Besides this major contribution, we share insight from the challenges we faced while curating this dataset. Based on these insights, we propose a novel and efficient labeling procedure that can be applied to any dataset. Our analysis examines its limitations and inductive biases, and based on these findings, we propose metrics to mitigate those limitations. Though our primary focus lies on visual similarity, the methodologies we present have broader applications for discovering and evaluating perceptual similarity across various domains.
****
Ego-Only: Egocentric Action Detection without Exocentric Transferring
Huiyu Wang, Mitesh Kumar Singh, Lorenzo Torresani; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5250-5261
We present Ego-Only, the first approach that enables state-of-the-art action detection on egocentric (first-person) videos without any form of exocentric (third-person) transferring. Despite the content and appearance gap separating the two domains, large-scale exocentric transferring has been the default choice for egocentric action detection. This is because prior works found that egocentric models are difficult to train from scratch and that transferring from exocentric representations leads to improved accuracy. However, in this paper, we revisit this common belief. Motivated by the large gap separating the two domains, we propose a strategy that enables effective training of egocentric models without exocentric transferring. Our Ego-Only approach is simple. It trains the video representation with a masked autoencoder finetuned for temporal segmentation. The learned features are then fed to an off-the-shelf temporal action localization method to detect actions. We find that this renders exocentric transferring unnecessary by showing remarkably strong results achieved by this simple Ego-Only approach on three established egocentric video datasets: Ego4D, EPIC-Kitchens-100, and Charades-Ego. On both action detection and action recognition, Ego-Only outperforms previous best exocentric transferring methods that use orders of magnitude more labels. Ego-Only sets new state-of-the-art results on these datasets and benchmarks without exocentric data.
****
CoinSeg: Contrast Inter- and Intra- Class Representations for Incremental Segmentation
Zekang Zhang, Guangyu Gao, Jianbo Jiao, Chi Harold Liu, Yunchao Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8

43-853
Class incremental semantic segmentation aims to strike a balance between the model's stability and plasticity by maintaining old knowledge while adapting to new concepts. However, most state-of-the-art methods use the freeze strategy for stability, which compromises the model's plasticity. In contrast, releasing parameter training for plasticity could lead to the best performance for all categories, but this requires discriminative feature representation. Therefore, we prioritize the model's plasticity and propose the Contrast inter- and intra-class representations for Incremental Segmentation (CoinSeg), which pursue discriminative representations for flexible parameter tuning. Inspired by the Gaussian mixture model that samples from a mixture of Gaussian distributions, CoinSeg emphasizes intra-class diversity
with multiple contrastive representation centroids. Specifically, we use mask proposals to identify regions with strong objectness that are likely to be diverse instances/centroids of a category. These mask proposals are then used for contrastive representations to reinforce intra-class diversity. Meanwhile, to avoid bias from intra-class diversity, we also apply category-level pseudo-labels to enhance category-level consistency and inter-category diversity. Additionally, CoinSeg ensures the model's stability and alleviates forgetting through a specific flexible tuning strategy. We validate CoinSeg on Pascal VOC 2012 and ADE20K datasets with multiple incremental scenarios and achieve superior results compared to previous state-of-the-art methods, especially in more challenging and realistic long-term scenarios.
****
Multi-View Active Fine-Grained Visual Recognition
Ruoyi Du, Wenqing Yu, Heqing Wang, Ting-En Lin, Dongliang Chang, Zhanyu Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1568-1578
Despite the remarkable progress of Fine-grained visual classification (FGVC) with years of history, it is still limited to recognizing 2 images. Recognizing objects in the physical world (i.e., 3D environment) poses a unique challenge -- discriminative information is not only present in visible local regions but also in other unseen views. Therefore, in addition to finding the distinguishable part from the current view, efficient and accurate recognition requires inferring the critical perspective with minimal glances. E.g., a person might recognize a "Ford sedan" with a glance at its side and then know that looking at the front can help tell which model it is. In this paper, towards FGVC in the real physical world, we put forward the problem of multi-view active fine-grained visual recognition (MAFR) and complete this study in three steps: (i) a multi-view, fine-grained vehicle dataset is collected as the testbed, (ii) a pilot experiment is designed to validate the need and research value of MAFR, (iii) a policy-gradient-based framework along with a dynamic exiting strategy is proposed to achieve efficient recognition with active view selection. Our comprehensive experiments demonstrate that the proposed method outperforms previous multi-view recognition works and can extend existing state-of-the-art FGVC methods and advanced neural networks to become FGVC experts in the 3D environment.
****
Part-Aware Transformer for Generalizable Person Re-identification
Hao Ni, Yuke Li, Lianli Gao, Heng Tao Shen, Jingkuan Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11280-11289
Domain generalization person re-identification (DG ReID) aims to train a model on source domains and generalize well on unseen domains. Vision Transformer usually yields better generalization ability than common CNN networks under distribution shifts. However, Transformer-based ReID models inevitably overfit to domain-specific biases due to the supervised learning strategy on the source domain. We observe that while the global images of different IDs should have different features, their similar local parts (e.g., black backpack) are not bounded by this constraint. Motivated by this, we propose a pure Transformer model (termed Part-aware Transformer) for DG-ReID by designing a proxy task, named Cross-ID Similarity Learning (CSL), to mine local visual information shared by different IDs. Th

is proxy task allows the model to learn generic features because it only cares about the visual similarity of the parts regardless of the ID labels, thus alleviating the side effect of domain-specific biases. Based on the local similarity obtained in CSL, a Part-guided Self-Distillation (PSD) is proposed to further improve the generalization of global features. Our method achieves state-of-the-art performance under most DG ReID settings. The code is available at https://github.com/liyuke65535/Part-Aware-Transformer.
****
Variational Causal Inference Network for Explanatory Visual Question Answering
Dizhan Xue, Shengsheng Qian, Changsheng Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2515-2525
Explanatory Visual Question Answering (EVQA) is a recently proposed multimodal reasoning task that requires answering visual questions and generating multimodal explanations for the reasoning processes. Unlike traditional Visual Question Answering (VQA) which focuses solely on answering, EVQA aims to provide user-friendly explanations to enhance the explainability and credibility of reasoning models. However, existing EVQA methods typically predict the answer and explanation separately, which ignores the causal correlation between them. Moreover, they neglect the complex relationships among question words, visual regions, and explanation tokens. To address these issues, we propose a Variational Causal Inference Network (VCIN) that establishes the causal correlation between predicted answers and explanations, and captures cross-modal relationships to generate rational explanations. First, we utilize a vision-and-language pretrained model to extract visual features and question features. Secondly, we propose a multimodal explanation gating transformer that constructs cross-modal relationships and generates rational explanations. Finally, we propose a variational causal inference to establish the target causal structure and predict the answers. Comprehensive experiments demonstrate the superiority of VCIN over state-of-the-art EVQA methods.
****
Improving Representation Learning for Histopathologic Images with Cluster Constraints
Weiyi Wu, Chongyang Gao, Joseph DiPalma, Soroush Vosoughi, Saeed Hassanpour; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21404-21414
Recent advances in whole-slide image (WSI) scanners and computational capabilities have significantly propelled the application of artificial intelligence in histopathology slide analysis. While these strides are promising, current supervised learning approaches for WSI analysis come with the challenge of exhaustively labeling high-resolution slides--a process that is both labor-intensive and time-consuming. In contrast, self-supervised learning (SSL) pretraining strategies are emerging as a viable alternative, given that they don't rely on explicit data annotations. These SSL strategies are quickly bridging the performance disparity with their supervised counterparts. In this context, we introduce an SSL framework. This framework aims for transferable representation learning and semantically meaningful clustering by synergizing invariance loss and clustering loss in WSI analysis. Notably, our approach outperforms common SSL methods in downstream classification and clustering tasks, as evidenced by tests on the Camelyon16 and a pancreatic cancer dataset.
****
Blending-NeRF: Text-Driven Localized Editing in Neural Radiance Fields
Hyeonseop Song, Seokhun Choi, Hoseok Do, Chul Lee, Taehyeong Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14383-14393
Text-driven localized editing of 3D objects is particularly difficult as locally mixing the original 3D object with the intended new object and style effects without distorting the object's form is not a straightforward process. To address this issue, we propose a novel NeRF-based model, Blending-NeRF, which consists of two NeRF networks: pretrained NeRF and editable NeRF. Additionally, we introduce new blending operations that allow Blending-NeRF to properly edit target regions which are localized by text. By using a pretrained vision-language aligned m

odel, CLIP, we guide Blending-NeRF to add new objects with varying colors and densities, modify textures, and remove parts of the original object. Our extensive experiments demonstrate that Blending-NeRF produces naturally and locally edited 3D objects from various text prompts.
****
Panoramas from Photons
Sacha Jungerman, Atul Ingle, Mohit Gupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10626-10636
Scene reconstruction in the presence of high-speed motion and low illumination is important in many applications such as augmented and virtual reality, drone navigation, and autonomous robotics. Traditional motion estimation techniques fail in such conditions, suffering from too much blur in the presence of high-speed motion and strong noise in low-light conditions. Single-photon cameras have recently emerged as a promising technology capable of capturing hundreds of thousands of photon frames per second thanks to their high speed and extreme sensitivity. Unfortunately, traditional computer vision techniques are not well suited for dealing with the binary-valued photon data captured by these cameras because these are corrupted by extreme Poisson noise. Here we present a method capable of estimating extreme scene motion under challenging conditions, such as low light or high dynamic range, from a sequence of high-speed image frames such as those captured by a single-photon camera. Our method relies on iteratively improving a motion estimate by grouping and aggregating frames after-the-fact, in a stratified manner. We demonstrate the creation of high-quality panoramas under fast motion and extremely low light, and super-resolution results using a custom single-photon camera prototype.
****
Global Adaptation Meets Local Generalization: Unsupervised Domain Adaptation for 3D Human Pose Estimation
Wenhao Chai, Zhongyu Jiang, Jenq-Neng Hwang, Gaoang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14655-14665
When applying a pre-trained 2D-to-3D Human Pose lifting model to a target unseen dataset, a large performance degradation is commonly encountered due to domain shift issues. We observe that the degradation is caused by two factors: 1) the large distribution gap over global positions of poses between the source and target datasets due to variant camera parameters and settings, and 2) the deficient diversity of local structures of poses in the training. To this end, we combine global adaptation and local generalization in PoseDA, a simple yet effective framework of unsupervised domain adaptation for 3D human pose estimation. Specifically, global adaptation aims to align global positions of poses from source domain to target domain with a proposed global position alignment (GPA) module. This module brings significant performance improvement without introducing additional learnable parameters. In addition, we propose local pose augmentation (LPA) to enhance the diversity of 3D poses following an adversarial training scheme consisting of 1) a augmentation generator that generates the parameters of pre-defined pose transformations and 2) an anchor discriminator to ensure the reality and quality of the augmented data. Our approach can be applicable to almost all 2D-3D lifting models. PoseDA achieves 61.3 mm of MPJPE on MPI-INF-3DHP under a cross-dataset evaluation setup, improving upon the previous state-of-the-art method by 10.2%.
****
Learning Neural Implicit Surfaces with Object-Aware Radiance Fields
Yiheng Zhang, Zhaofan Qiu, Yingwei Pan, Ting Yao, Tao Mei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17893-17902
Recent progress on multi-view 3D object reconstruction has featured neural implicit surfaces via learning high-fidelity radiance fields. However, most approaches hinge on the visual hull derived from cost-expensive silhouette masks to obtain object surfaces. In this paper, we propose a novel Object-aware Radiance Fields (ORF) to automatically learn an object-aware geometry reconstruction. The geometric correspondences between multi-view 2D object regions and 3D implicit/explicit object surfaces are additionally exploited to boost the learning of object s

urfaces. Technically, a critical transparency discriminator is designed to distinguish the object-intersected and object-bypassed rays based on the estimated 2D object regions, leading to 3D implicit object surfaces. Such implicit surfaces can be directly converted into explicit object surfaces (e.g., meshes) via marching cubes. Then, we build the geometric correspondence between 2D planes and 3D meshes by rasterization, and project the estimated object regions into 3D explicit object surfaces by aggregating the object information across multiple views. The aggregated object information in 3D explicit object surfaces is further reprojected back to 2D planes, aiming to update 2D object regions and enforce them to be multi-view consistent. Extensive experiments on DTU and BlendedMVS verify the capability of ORF to produce comparable surfaces against the state-of-the-art models that demand silhouette masks.
****
PADCLIP: Pseudo-labeling with Adaptive Debiasing in CLIP for Unsupervised Domain Adaptation
Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, Chen-Nee Chuah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16155-16165
Traditional Unsupervised Domain Adaptation (UDA) leverages the labeled source domain to tackle the learning tasks on the unlabeled target domain. It can be more challenging when a large domain gap exists between the source and the target domain. A more practical setting is to utilize a large-scale pre-trained model to fill the domain gap. For example, CLIP shows promising zero-shot generalizability to bridge the gap. However, after applying traditional fine-tuning to specifically adjust CLIP on a target domain, CLIP suffers from catastrophic forgetting issues where the new domain knowledge can quickly override CLIP's pre-trained knowledge and decreases the accuracy by half. We propose Catastrophic Forgetting Measurement (CFM) to adjust the learning rate to avoid excessive training (thus mitigating the catastrophic forgetting issue). We then utilize CLIP's zero-shot prediction to formulate a Pseudo-labeling setting with Adaptive Debiasing in CLIP (PADCLIP) by adjusting causal inference with our momentum and CFM. Our PADCLIP allows end-to-end training on source and target domains without extra overhead, and we achieved the best results on four public datasets, with a significant improvement (+18.5% accuracy) on DomainNet.
****
Causal-DFQ: Causality Guided Data-Free Network Quantization
Yuzhang Shang, Bingxin Xu, Gaowen Liu, Ramana Rao Kompella, Yan Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17437-17446
Model quantization, which aims to compress deep neural networks and accelerate inference speed, has greatly facilitated the development of cumbersome models on mobile and edge devices. There is a common assumption in quantization methods from prior works that training data is available. In practice, however, this assumption cannot always be fulfilled due to reasons of privacy and security, rendering these methods inapplicable in real-life situations.
Thus, data-free network quantization has recently received significant attention in neural network compression. Causal reasoning provides an intuitive way to model casual relationships to eliminate data-driven correlations, making causality an essential component of analyzing data-free problems. However, causal formulations of data-free quantization are inadequate in the literature. To bridge this gap, we construct a causal graph to model the data generation and discrepancy reduction between the pre-trained and quantized models. Inspired by the causal understanding, we propose the Causality-guided Data-free Network Quantization method, Causal-DFQ, to eliminate the reliance on data via approaching an equilibrium of causality-driven intervened distributions. Specifically, we design a content-style-decoupled generator, synthesizing images conditioned on the relevant and irrelevant factors; then we propose a discrepancy reduction loss to align the intervened distributions of the pre-trained and quantized models. It is worth noting that our work is the first attempt towards introducing causality to data-free quantization problem. Extensive experiments demonstrate the efficacy of Causal

-DFQ.
****
Enhancing Generalization of Universal Adversarial Perturbation through Gradient Aggregation
Xuannan Liu, Yaoyao Zhong, Yuhang Zhang, Lixiong Qin, Weihong Deng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4435-4444
Deep neural networks are vulnerable to universal adversarial perturbation (UAP), an instance-agnostic perturbation capable of fooling the target model for most samples. Compared to instance-specific adversarial examples, UAP is more challenging as it needs to generalize across various samples and models. In this paper, we examine the serious dilemma of UAP generation methods from a generalization perspective -- the gradient vanishing problem using small-batch stochastic gradient optimization and the local optima problem using large-batch optimization. To address these problems, we propose a simple and effective method called Stochastic Gradient Aggregation (SGA), which alleviates the gradient vanishing and escapes from poor local optima at the same time. Specifically, SGA employs the small-batch training to perform multiple iterations of inner pre-search. Then, all the inner gradients are aggregated as a one-step gradient estimation to enhance the gradient stability and reduce quantization errors. Extensive experiments on the standard ImageNet dataset demonstrate that our method significantly enhances the generalization ability of UAP and outperforms other state-of-the-art methods. The code is available at https://github.com/liuxuannan/Stochastic-Gradient-Aggregation.
****
CancerUniT: Towards a Single Unified Model for Effective Detection, Segmentation, and Diagnosis of Eight Major Cancers Using a Large Collection of CT Scans
Jieneng Chen, Yingda Xia, Jiawen Yao, Ke Yan, Jianpeng Zhang, Le Lu, Fakai Wang, Bo Zhou, Mingyan Qiu, Qihang Yu, Mingze Yuan, Wei Fang, Yuxing Tang, Minfeng Xu, Jian Zhou, Yuqian Zhao, Qifeng Wang, Xianghua Ye, Xiaoli Yin, Yu Shi, Xin Chen, Jingren Zhou, Alan Yuille, Zaiyi Liu, Ling Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21327-21338
Human readers or radiologists routinely perform full-body multi-organ multi-disease detection and diagnosis in clinical practice, while most medical AI systems are built to focus on single organs with a narrow list of a few diseases. This might severely limit AI's clinical adoption. A certain number of AI models need to be assembled non-trivially to match the diagnostic process of a human reading a CT scan. In this paper, we construct a Unified Tumor Transformer (CancerUniT) model to jointly detect tumor existence & location and diagnose tumor characteristics for eight major cancers in CT scans. CancerUniT is a query-based Mask Transformer model with the output of multi-tumor prediction. We decouple the object queries into organ queries, tumor detection queries and tumor diagnosis queries, and further establish hierarchical relationships among the three groups. This clinically-inspired architecture effectively assists inter- and intra-organ representation learning of tumors and facilitates the resolution of these complex, anatomically related multi-organ cancer image reading tasks. CancerUniT is trained end-to-end using curated large-scale CT images of 10,042 patients including eight major types of cancers and occurring non-cancer tumors (all are pathology-confirmed with 3D tumor masks annotated by radiologists). On the test set of 631 patients, CancerUniT has demonstrated strong performance under a set of clinically relevant evaluation metrics, substantially outperforming both multi-disease methods and an assembly of eight single-organ expert models in tumor detection, segmentation, and diagnosis. This moves one step closer towards a universal high performance cancer screening tool.
****
Dual Meta-Learning with Longitudinally Consistent Regularization for One-Shot Brain Tissue Segmentation Across the Human Lifespan
Yongheng Sun, Fan Wang, Jun Shu, Haifeng Wang, Li Wang, Deyu Meng, Chunfeng Lian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21118-21128

Brain tissue segmentation is essential for neuroscience and clinical studies. However, segmentation on longitudinal data is challenging due to dynamic brain changes across the lifespan. Previous researches mainly focus on self-supervision with regularizations and will lose longitudinal generalization when fine-tuning on a specific age group. In this paper, we propose a dual meta-learning paradigm to learn longitudinally consistent representations and persist when fine-tuning. Specifically, we learn a plug-and-play feature extractor to extract longitudinal-consistent anatomical representations by meta-feature learning and a well-initialized task head for fine-tuning by meta-initialization learning. Besides, two class-aware regularizations are proposed to encourage longitudinal consistency. Experimental results on the iSeg2019 and ADNI datasets demonstrate the effectiveness of our method.
****

DeFormer: Integrating Transformers with Deformable Models for 3D Shape Abstraction from a Single Image

Di Liu, Xiang Yu, Meng Ye, Qilong Zhangli, Zhuowei Li, Zhixing Zhang, Dimitris N. Metaxas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14236-14246

Explicit 3D shape abstraction from a single 2D image is a long-standing problem in computer vision and graphics. By leveraging a set of primitives to represent the target shape, recent methods have achieved promising results. However, these methods either use a relatively larger number of primitives or lack geometric flexibility due to the low-dimensional expressibility of the primitives. In this paper, we propose a novel bi-channel Transformer architecture, integrated with parameterized deformable models, termed DeFormer, to simultaneously estimate global and local deformations. In this way, DeFormer can abstract complex object shapes while using a small number of primitives which offer a broader geometry coverage and finer details. Then, we introduce a force-driven dynamic fitting and a cycle-consistent re-projection loss to optimize the primitive parameters. Extensive experiments on ShapeNet across various settings show that DeFormer achieves better reconstruction accuracy over the state-of-the-art, and visualizes with consistent semantic correspondences for improved interpretability.
****

Parallel Attention Interaction Network for Few-Shot Skeleton-Based Action Recognition

Xingyu Liu, Sanping Zhou, Le Wang, Gang Hua; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1379-1388

Learning discriminative features from very few labeled samples to identify novel classes has received increasing attention in skeleton-based action recognition. Existing works aim to learn action-specific embeddings by exploiting either intra-skeleton or inter-skeleton spatial associations, which may lead to less discriminative representations. To address these issues, we propose a novel Parallel Attention Interaction Network (PAINet) that incorporates two complementary branches to strengthen the match by inter-skeleton and intra-skeleton correlation. Specifically, a topology encoding module utilizing topology and physical information is proposed to enhance the modeling of interactive parts and joint pairs in both branches. In the Cross Spatial Alignment branch, we employ a spatial cross-attention module to establish joint associations across sequences, and a directional Average Symmetric Surface Metric is introduced to locate the closest temporal similarity. In parallel, the Cross Temporal Alignment branch incorporates a spatial self-attention module to aggregate spatial context within sequences as well as applies the temporal cross-attention network to correct misalignment temporally and calculate similarity. Extensive experiments on three skeleton benchmarks, namely NTU-T, NTU-S, and Kinetics, demonstrate the superiority of our framework and consistently outperform state-of-the-art methods.
****

Cross-view Semantic Alignment for Livestreaming Product Recognition

Wenjie Yang, Yiyi Chen, Yan Li, Yanhua Cheng, Xudong Liu, Quan Chen, Han Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13404-13413

Live commerce is the act of selling products online through livestreaming. The c
ustomer's diverse demands for online products introduces more challenges to Live
streaming Product Recognition. Previous works are either focus on fashion clothi
ng data or subject to single-modal input, thus inconsistent with the real-world
scenario where multimodal data from various categories are present. In this pape
r, we contribute LPR4M, a large-scale multimodal dataset that covers 34 categori
es, comprises 3 modalities (image, video, and text), and is 50 times larger than
 the largest publicly available dataset. In addition, LPR4M contains diverse vid
eos and noise modality pair while also having a long-tailed distribution, resemb
ling real-world problems. Moreover, a cRoss-vIew semantiC alignmEnt (RICE) model
 is proposed to learn discriminative instance features from the two views (image
 and video) of products via instance-level contrastive learning as well as cross
-view patch-level feature propagation. A novel Patch Feature Reconstruction loss
 is proposed to penalize the semantic misalignment between the cross-view patche
s. Extensive ablation studies demonstrate the effectiveness of RICE and provide
insights into the importance of dataset diversity and expressivity.
****
Continuously Masked Transformer for Image Inpainting
Keunsoo Ko, Chang-Su Kim; Proceedings of the IEEE/CVF International Conference o
n Computer Vision (ICCV), 2023, pp. 13169-13178
A novel continuous-mask-aware transformer for image inpainting, called CMT, is p
roposed in this paper, which uses a continuous mask to represent the amounts of
errors in tokens. First, we initialize a mask and use it during the self-attenti
on. To facilitate the masked self-attention, we also introduce the notion of ove
rlapping tokens. Second, we update the mask by modeling the error propagation du
ring the masked self-attention. Through several masked self-attention and mask u
pdate (MSAU) layers, we predict initial inpainting results. Finally, we refine t
he initial results to reconstruct a more faithful image. Experimental results on
 multiple datasets show that the proposed CMT algorithm outperforms existing alg
orithms significantly. The source codes are available at https://github.com/keun
soo-ko/CMT.
****
Vanishing Point Estimation in Uncalibrated Images with Prior Gravity Direction
Rémi Pautrat, Shaohui Liu, Petr Hruby, Marc Pollefeys, Daniel Barath; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
14118-14127
We tackle the problem of estimating a Manhattan frame, i.e. three orthogonal van
ishing points, and the unknown focal length of the camera, leveraging a prior ve
rtical direction. The direction can come from an Inertial Measurement Unit that
is a standard component of recent consumer devices, e.g., smartphones. We provid
e an exhaustive analysis of minimal line configurations and derive two new 2-lin
e solvers, one of which does not suffer from singularities affecting existing so
lvers. Additionally, we design a new non-minimal method, running on an arbitrary
 number of lines, to boost the performance in local optimization. Combining all
solvers in a hybrid robust estimator, our method achieves increased accuracy eve
n with a rough prior. Experiments on synthetic and real-world datasets demonstra
te the superior accuracy of our method compared to the state of the art, while h
aving comparable runtimes. We further demonstrate the applicability of our solve
rs for relative rotation estimation. The code is available at https://github.com
/cvg/VP-Estimation-with-Prior-Gravity.
****
Learn TAROT with MENTOR: A Meta-Learned Self-Supervised Approach for Trajectory
Prediction
Mozhgan Pourkeshavarz, Changhe Chen, Amir Rasouli; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 8384-8393
Predicting diverse yet admissible trajectories that adhere to the map constraint
s is challenging. Graph-based scene encoders have been proven effective for pres
erving local structures of maps by defining lane-level connections. However, suc
h encoders do not capture more complex patterns emerging from long-range heterog
eneous connections between nonadjacent interacting lanes. To this end, we shed n

ew light on learning common driving patterns by introducing meTA ROad paTh (TARO
T) to formulate combinations of various relations between lanes on the road topo
logy. Intuitively, this can be viewed as finding feasible routes. Furthermore, w
e propose MEta-road NeTwORk (MENTOR) that helps trajectory prediction by providi
ng it with TAROT as navigation tips. More specifically, 1) we define TAROT predi
ction as a novel self-supervised proxy task to identify the complex heterogeneou
s structure of the map. 2) For typical driving actions, we establish several TAR
OTs that result in multiple Heterogeneous Structure Learning (HSL) tasks. These
tasks are used in MENTOR, which performs meta-learning by simultaneously predict
ing trajectories along with proxy tasks, identifying an optimal combination of t
hem, and automatically balancing them to improve the primary task. We show that
our model achieves state-of-the-art performance on the Argoverse dataset, especi
ally on diversity and admissibility metrics, achieving up to 20% improvements in
 challenging scenarios. We further investigate the contribution of proposed modu
les in ablation studies.
****
MatrixVT: Efficient Multi-Camera to BEV Transformation for 3D Perception
Hongyu Zhou, Zheng Ge, Zeming Li, Xiangyu Zhang; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 8548-8557
This paper proposes an efficient multi-camera to Bird's-Eye-View (BEV) view tran
sformation method for 3D perception, dubbed MatrixVT.
 Existing view transformers either suffer from poor transformation efficiency or
 rely on device-specific operators, hindering the broad application of BEV model
s. In contrast, our method generates BEV features efficiently with only convolut
ions and matrix multiplications (MatMul). Specifically, we propose describing th
e BEV feature as the MatMul of image feature and a sparse Feature Transporting M
atrix (FTM). A Prime Extraction module is then introduced to compress the dimens
ion of image features and reduce FTM's sparsity. Moreover, we propose the Ring &
 Ray Decomposition to replace the FTM with two matrices and reformulate our pipe
line to reduce calculation further. Compared to existing methods, MatrixVT enjoy
s a faster speed and less memory footprint while remaining deploy-friendly. Exte
nsive experiments on the nuScenes benchmark demonstrate that our method is highl
y efficient but obtains results on par with the SOTA method in object detection
and map segmentation tasks. Code will be available.
****
Local and Global Logit Adjustments for Long-Tailed Learning
Yingfan Tao, Jingna Sun, Hao Yang, Li Chen, Xu Wang, Wenming Yang, Daniel Du, Mi
n Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision
 (ICCV), 2023, pp. 11783-11792
Multi-expert ensemble models for long-tailed learning typically either learn div
erse generalists from the whole dataset or aggregate specialists on different su
bsets. However, the former is insufficient for tail classes due to the high imba
lance factor of the entire dataset, while the latter may bring ambiguity in pred
icting unseen classes. To address these issues, we propose a novel Local and Glo
bal Logit Adjustments (LGLA) method that learns experts with full data covering
all classes and enlarges the discrepancy among them by elaborated logit adjustme
nts. LGLA consists of two core components: a Class-aware Logit Adjustment (CLA)
strategy and an Adaptive Angular Weighted (AAW) loss. The CLA strategy trains mu
ltiple experts which excel at each subset using the Local Logit Adjustment (LLA)
. It also trains one expert specializing in an inversely long-tailed distributio
n through Global Logit Adjustment (GLA). Moreover, the AAW loss adopts adaptive
hard sample mining with respect to different experts to further improve accuracy
. Extensive experiments on popular long-tailed benchmarks manifest the superiori
ty of LGLA over the SOTA methods.
****
Active Self-Supervised Learning: A Few Low-Cost Relationships Are All You Need
Vivien Cabannes, Leon Bottou, Yann Lecun, Randall Balestriero; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16274-1
6283
Self-Supervised Learning (SSL) has emerged as the solution of choice to learn tr

ansferable representations from unlabeled data. However, SSL requires to build samples that are known to be semantically akin, i.e. positive views.
Requiring such knowledge is the main limitation of SSL and is often tackled by ad-hoc strategies e.g. applying known data-augmentations to the same input.
In this work, we generalize and formalize this principle through Positive Active Learning (PAL) where an oracle queries semantic relationships between samples.
PAL achieves three main objectives. First, it is a theoretically grounded learning framework that encapsulates standard SSL but also supervised and semi-supervised learning depending on the employed oracle.
Second, it provides a consistent algorithm to embed a priori knowledge, e.g. some observed labels, into any SSL losses without any change in the training pipeline.
Third, it provides a proper active learning framework yielding low-cost solutions to annotate datasets, arguably bringing the gap between theory and practice of active learning that is based on simple-to-answer-by-non-experts queries of semantic relationships between inputs.
****
Wasserstein Expansible Variational Autoencoder for Discriminative and Generative Continual Learning
Fei Ye, Adrian G. Bors; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18665-18675
Task-Free Continual Learning (TFCL) represents a challenging learning paradigm where a model is trained on the non-stationary data distributions without any knowledge of the task information, thus representing a more practical approach. Despite promising achievements by the Variational Autoencoder (VAE) mixtures in continual learning, such methods ignore the redundancy among the probabilistic representations of their components when performing model expansion, leading to mixture components learning similar tasks. This paper proposes the Wasserstein Expansible Variational Autoencoder (WEVAE), which evaluates the statistical similarity between the probabilistic representation of new data and that represented by each mixture component and then uses it for deciding when to expand the model. Such a mechanism can avoid unnecessary model expansion while ensuring the knowledge diversity among the trained components. In addition, we propose an energy-based sample selection approach that assigns high energies to novel samples and low energies to the samples which are similar to the model's knowledge. Extensive empirical studies on both supervised and unsupervised benchmark tasks demonstrate that our model outperforms all competing methods. The code is available at https://github.com/dtuzi123/WEVAE/.
****
Sensitivity-Aware Visual Parameter-Efficient Fine-Tuning
Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, Bohan Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11825-11835
Visual Parameter-Efficient Fine-Tuning (PEFT) has become a powerful alternative for full fine-tuning so as to adapt pre-trained vision models to downstream tasks, which only tunes a small number of parameters while freezing the vast majority ones to ease storage burden and optimization difficulty. However, existing PEFT methods introduce trainable parameters to the same positions across different tasks depending solely on human heuristics and neglect the domain gaps. To this end, we study where to introduce and how to allocate trainable parameters by proposing a novel Sensitivity-aware visual Parameter-efficient fine-Tuning (SPT) scheme, which adaptively allocates trainable parameters to task-specific important positions given a desired tunable parameter budget. Specifically, our SPT first quickly identifies the sensitive parameters that require tuning for a given task in a data-dependent way. Next, our SPT further boosts the representational capability for the weight matrices whose number of sensitive parameters exceeds a pre-defined threshold by utilizing existing structured tuning methods, e.g., LoRA or Adapter, to replace directly tuning the selected sensitive parameters (unstructured tuning) under the budget. Extensive experiments on a wide range of downstream recognition tasks show that our SPT is complementary to the existing PEFT

methods and largely boosts their performance, e.g., SPT improves Adapter with supervised pre-trained ViT-B/16 backbone by 4.2% and 1.4% mean Top-1 accuracy, reaching SOTA performance on FGVC and VTAB-1k benchmarks, respectively. Source code is at https://github.com/ziplab/SPT.

****

Label-Free Event-based Object Recognition via Joint Learning with Image Reconstruction from Events

Hoonhee Cho, Hyeonseong Kim, Yujeong Chae, Kuk-Jin Yoon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19866-19877

Recognizing objects from sparse and noisy events becomes extremely difficult when paired images and category labels do not exist. In this paper, we study label-free event-based object recognition where category labels and paired images are not available. To this end, we propose a joint formulation of object recognition and image reconstruction in a complementary manner. Our method first reconstructs images from events and performs object recognition through Contrastive Language-Image Pre-training (CLIP), enabling better recognition through a rich context of images. Since the category information is essential in reconstructing images, we propose category-guided attraction loss and category-agnostic repulsion loss to bridge the textual features of predicted categories and the visual features of reconstructed images using CLIP. Moreover, we introduce a reliable data sampling strategy and local-global reconstruction consistency to boost joint learning of two tasks. To enhance the accuracy of prediction and quality of reconstruction, we also propose a prototype-based approach using unpaired images. Extensive experiments demonstrate the superiority of our method and its extensibility for zero-shot object recognition. Our project code is available at https://github.com/Chohoonhee/Ev-LaFOR.

****

Gloss-Free Sign Language Translation: Improving from Visual-Language Pretraining

Benjia Zhou, Zhigang Chen, Albert Clapés, Jun Wan, Yanyan Liang, Sergio Escalera, Zhen Lei, Du Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20871-20881

Sign Language Translation (SLT) is a challenging task due to its cross-domain nature, involving the translation of visual-gestural language to text. Many previous methods employ an intermediate representation,i.e., gloss sequences, to facilitate SLT, thus transforming it into a two-stage task of sign language recognition (SLR) followed by sign language translation (SLT). However, the scarcity of gloss-annotated sign language data, combined with the information bottleneck in the mid-level gloss representation, has hindered the further development of the SLT task. To address this challenge, we propose a novel Gloss-Free SLT base on Visual-Language Pretraining (GFSLT-VLP), which improves SLT by inheriting language-oriented prior knowledge from pre-trained models, without any gloss annotation assistance. Our approach involves two stages: (i) integrating Contrastive Language-Image Pre-training (CLIP) with masked self-supervised learning to create pre-tasks that bridge the semantic gap between visual and textual representations and restore masked sentences, and (ii) constructing an end-to-end architecture with an encoder-decoder-like structure that inherits the parameters of the pre-trained Visual Encoder and Text Decoder from the first stage. The seamless combination of these novel designs forms a robust sign language representation and significantly improves gloss-free sign language translation. In particular, we have achieved unprecedented improvements in terms of BLEU-4 score on the PHOENIX14T dataset (>=+5) and the CSL-Daily dataset (>=+3) compared to state-of-the-art gloss-free SLT methods. Furthermore, our approach also achieves competitive results on the PHOENIX14T dataset when compared with most of the gloss-based methods.

****

Weakly-supervised 3D Pose Transfer with Keypoints

Jinnan Chen, Chen Li, Gim Hee Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15156-15165

The main challenges of 3D pose transfer are: 1) Lack of paired training data with different characters performing the same pose; 2) Disentangling pose and shape information from the target mesh; 3) Difficulty in applying to meshes with diff

erent topologies. We thus propose a novel weakly-supervised keypoint-based frame work to overcome these difficulties. Specifically, we use a topology-agnostic ke ypoint detector with inverse kinematics to compute transformations between the s ource and target meshes. Our method only requires supervision on the keypoints, can be applied to meshes with different topologies and is shape-invariant for th e target which allows extraction of pose-only information from the target meshes without transferring shape information. We further design a cycle reconstructio n to perform self-supervised pose transfer without the need for ground truth def ormed mesh with the same pose and shape as the target and source, respectively. We evaluate our approach on benchmark human and animal datasets, where we achiev e superior performance compared to the state-of-the-art unsupervised approaches and even comparable performance with the fully supervised approaches. We test on the more challenging Mixamo dataset to verify our approach's ability in handlin g meshes with different topologies and complex clothes. Cross-dataset evaluation further shows the strong generalization ability of our approach. Our code will be open-sourced upon paper acceptance.
****
Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement
Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, Peng Gao ; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) , 2023, pp. 2605-2615
The popularity of Contrastive Language-Image Pre-training (CLIP) has propelled i ts application to diverse downstream vision tasks. To improve its capacity on do wnstream tasks, few-shot learning has become a widely-adopted technique. However , existing methods either exhibit limited performance or suffer from excessive l earnable parameters. In this paper, we propose APE, an Adaptive Prior rEfinement method for CLIP's pre-trained knowledge, which achieves superior accuracy with high computational efficiency. Via a prior refinement module, we analyze the int er-class disparity in the downstream data and decouple the domain-specific knowl edge from the CLIP-extracted cache model. On top of that, we introduce two model variants, a training-free APE and a training-required APE-T. We explore the tri lateral affinities between the test image, prior cache model, and textual repres entations, and only enable a lightweight category-residual module to be trained. For the average accuracy over 11 benchmarks, both APE and APE-T attain state-of -the-art and respectively outperform the second-best by +1.59% and +1.99% under 16 shots with x30 less learnable parameters.
****
EgoVLPv2: Egocentric Video-Language Pre-training with Fusion in the Backbone
Shraman Pramanick, Yale Song, Sayan Nag, Kevin Qinghong Lin, Hardik Shah, Mike Z heng Shou, Rama Chellappa, Pengchuan Zhang; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2023, pp. 5285-5297
Video-language pre-training (VLP) has become increasingly important due to its a bility to generalize to various vision and language tasks. However, existing ego centric VLP frameworks utilize separate video and language encoders and learn ta sk-specific cross-modal information only during fine-tuning, limiting the develo pment of a unified system. In this work, we introduce the second generation of e gocentric video-language pre-training (EgoVLPv2), a significant improvement from the previous generation, by incorporating cross-modal fusion directly into the video and language backbones. EgoVLPv2 learns strong video-text representation d uring pre-training and reuses the cross-modal attention modules to support diffe rent downstream tasks in a flexible and efficient manner, reducing fine-tuning c osts. Moreover, our proposed fusion in the backbone strategy is more lightweight and compute-efficient than stacking additional fusion-specific layers. Extensiv e experiments on a wide range of VL tasks demonstrate the effectiveness of EgoVL Pv2 by achieving consistent state-of-the-art performance over strong baselines a cross all downstream.
****
On the Effectiveness of Spectral Discriminators for Perceptual Quality Improveme nt
Xin Luo, Yunan Zhu, Shunxin Xu, Dong Liu; Proceedings of the IEEE/CVF Internatio

nal Conference on Computer Vision (ICCV), 2023, pp. 13243-13253
Several recent studies advocate the use of spectral discriminators, which evaluate the Fourier spectra of images for generative modeling. However, the effectiveness of the spectral discriminators is not well interpreted yet. We tackle this issue by examining the spectral discriminators in the context of perceptual image super-resolution (i.e., GAN-based SR), as SR image quality is susceptible to spectral changes. Our analyses reveal that the spectral discriminator indeed performs better than the ordinary (a.k.a. spatial) discriminator in identifying the differences in the high-frequency range; however, the spatial discriminator holds an advantage in the low-frequency range. Thus, we suggest that the spectral and spatial discriminators shall be used simultaneously. Moreover, we improve the spectral discriminators by first calculating the patch-wise Fourier spectrum and then aggregating the spectra by Transformer. We verify the effectiveness of the proposed method twofold. On the one hand, thanks to the additional spectral discriminator, our obtained SR images have their spectra better aligned to those of the real images, which leads to a better PD tradeoff. On the other hand, our ensembled discriminator predicts the perceptual quality more accurately, as evidenced in the no-reference image quality assessment task.
****
Shrinking Class Space for Enhanced Certainty in Semi-Supervised Learning
Lihe Yang, Zhen Zhao, Lei Qi, Yu Qiao, Yinghuan Shi, Hengshuang Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16187-16196
Semi-supervised learning is attracting blooming attention, due to its success in combining unlabeled data. To mitigate potentially incorrect pseudo labels, recent frameworks mostly set a fixed confidence threshold to discard uncertain samples. This practice ensures high-quality pseudo labels, but incurs a relatively low utilization of the whole unlabeled set. In this work, our key insight is that these uncertain samples can be turned into certain ones, as long as the confusion classes for the top-1 class are detected and removed. Invoked by this, we propose a novel method dubbed ShrinkMatch to learn uncertain samples. For each uncertain sample, it adaptively seeks a shrunk class space, which merely contains the original top-1 class, as well as remaining less likely classes. Since the confusion ones are removed in this space, the re-calculated top-1 confidence can satisfy the pre-defined threshold. We then impose a consistency regularization between a pair of strongly and weakly augmented samples in the shrunk space to strive for discriminative representations. Furthermore, considering the varied reliability among uncertain samples and the gradually improved model during training, we correspondingly design two reweighting principles for our uncertain loss. Our method exhibits impressive performance on widely adopted benchmarks. Code is available at https://github.com/LiheYoung/ShrinkMatch.
****
Deep Equilibrium Object Detection
Shuai Wang, Yao Teng, Limin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6296-6306
Query-based object detectors directly decode image features into object instances with a set of learnable queries. These query vectors are progressively refined to stable meaningful representations through a sequence of decoder layers, and then used to directly predict object locations and categories with simple FFN heads.
In this paper, we present a new query-based object detector (DEQDet) by designing a deep equilibrium decoder. Our DEQ decoder models the query vector refinement as the fixed point solving of an implicit layer and is equivalent to applying infinite steps of refinement.
To be more specific to object decoding, we use a two-step unrolled equilibrium equation to explicitly capture the query vector refinement. Accordingly, we are able to incorporate refinement awareness into the DEQ training with the inexact gradient back-propagation (RAG). In addition, to stabilize the training of our DEQDet and improve its generalization ability, we devise the deep supervision scheme on the optimization path of DEQ with refinement-aware perturbation (RAP).

Our experiments demonstrate DEQDet converges faster, consumes less memory, and achieves better results than the baseline counterpart (AdaMixer). In particular, our DEQDet with ResNet50 backbone and 300 queries achieves the 49.5 mAP and 33.0 APs on the MS COCO benchmark under 2x training scheme (24 epochs).
****
Diffusion-Based 3D Human Pose Estimation with Multi-Hypothesis Aggregation
Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, Wen Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14761-14771
In this paper, a novel Diffusion-based 3D Pose estimation (D3DP) method with Joint-wise reProjection-based Multi-hypothesis Aggregation (JPMA) is proposed for probabilistic 3D human pose estimation. On the one hand, D3DP generates multiple possible 3D pose hypotheses for a single 2D observation. It gradually diffuses the ground truth 3D poses to a random distribution, and learns a denoiser conditioned on 2D keypoints to recover the uncontaminated 3D poses. The proposed D3DP is compatible with existing 3D pose estimators and supports users to balance efficiency and accuracy during inference through two customizable parameters. On the other hand, JPMA is proposed to assemble multiple hypotheses generated by D3DP into a single 3D pose for practical use. It reprojects 3D pose hypotheses to the 2D camera plane, selects the best hypothesis joint-by-joint based on the reprojection errors, and combines the selected joints into the final pose. The proposed JPMA conducts aggregation at the joint level and makes use of the 2D prior information, both of which have been overlooked by previous approaches. Extensive experiments on Human3.6M and MPI-INF-3DHP datasets show that our method outperforms the state-of-the-art deterministic and probabilistic approaches by 1.5% and 8.9%, respectively. Code is available at https://github.com/paTRICK-swk/D3DP.
****
RPEFlow: Multimodal Fusion of RGB-PointCloud-Event for Joint Optical Flow and Scene Flow Estimation
Zhexiong Wan, Yuxin Mao, Jing Zhang, Yuchao Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10030-10040
Recently, the RGB images and point clouds fusion methods have been proposed to jointly estimate 2D optical flow and 3D scene flow. However, as both conventional RGB cameras and LiDAR sensors adopt a frame-based data acquisition mechanism, their performance is limited by the fixed low sampling rates, especially in highly-dynamic scenes. By contrast, the event camera can asynchronously capture the intensity changes with a very high temporal resolution, providing complementary dynamic information of the observed scenes. In this paper, we incorporate RGB images, Point clouds and Events for joint optical flow and scene flow estimation with our proposed multi-stage multimodal fusion model, RPEFlow. First, we present an attention fusion module with a cross-attention mechanism to implicitly explore the internal cross-modal correlation for 2D and 3D branches, respectively. Second, we introduce a mutual information regularization term to explicitly model the complementary information of three modalities for effective multimodal feature learning. We also contribute a new synthetic dataset to advocate further research. Experiments on both synthetic and real datasets show that our model outperforms the existing state-of-the-art by a wide margin. Code and dataset is available at https://npucvr.github.io/RPEFlow.
****
SMAUG: Sparse Masked Autoencoder for Efficient Video-Language Pre-Training
Yuanze Lin, Chen Wei, Huiyu Wang, Alan Yuille, Cihang Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2459-2469
Video-language pre-training is crucial for learning powerful multi-modal representation. However, it typically requires a massive amount of computation. In this paper, we develop SMAUG, an efficient pre-training framework for video-language models. The foundation component in SMAUG is masked autoencoders. Different from prior works which only mask textual inputs, our masking strategy considers both visual and textual modalities, providing a better cross-modal alignment and saving more pre-training costs. On top of that, we introduce a space-time token sparsification module, which leverages context information to further select only

"important" spatial regions and temporal frames for pre-training. Coupling all these designs allows our method to enjoy both competitive performances on text-to-video retrieval and video question answering tasks, and much less pre-training costs by 1.9x or more. For example, our SMAUG only needs 50 NVIDIA A6000 GPU hours for pre-training to attain competitive performances on these two video-language tasks across six popular benchmarks.
****
eP-ALM: Efficient Perceptual Augmentation of Language Models
Mustafa Shukor, Corentin Dancette, Matthieu Cord; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22056-22069
Large Language Models (LLMs) have so far impressed the world, with unprecedented capabilities that emerge in models at large scales. On the vision side, transformer models (i.e., ViT) are following the same trend, achieving the best performance on challenging benchmarks. With the abundance of such unimodal models, a natural question arises; do we need also to follow this trend to tackle multimodal tasks? In this work, we propose to rather direct effort to efficient adaptations of existing models, and propose to augment Language Models with perception. Existing approaches for adapting pretrained models for vision-language tasks still rely on several key components that hinder their efficiency. In particular, they still train a large number of parameters, rely on large multimodal pretraining, use encoders (e.g., CLIP) trained on huge image-text datasets, and add significant inference overhead. In addition, most of these approaches have focused on Zero-Shot and In Context Learning, with little to no effort on direct finetuning. We investigate the minimal computational effort needed to adapt unimodal models for multimodal tasks and propose a new challenging setup, alongside different approaches, that efficiently adapts unimodal pretrained models.
We show that by freezing more than 99% of total parameters, training only one linear projection layer, and prepending only one trainable token, our approach (dubbed eP-ALM) significantly outperforms other baselines on VQA and Captioning across Image, Video, and Audio modalities, following the proposed setup. The code is available here: https://github.com/mshukor/eP-ALM.
****
Multimodal Optimal Transport-based Co-Attention Transformer with Global Structure Consistency for Survival Prediction
Yingxue Xu, Hao Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21241-21251
Survival prediction is a complicated ordinal regression task that aims to predict the ranking risk of death, which generally benefits from the integration of histology and genomic data. Despite the progress in joint learning from pathology and genomics, existing methods still suffer from challenging issues: 1) Due to the large size of pathological images, it is difficult to effectively represent the gigapixel whole slide images (WSIs). 2) Interactions within tumor microenvironment (TME) in histology are essential for survival analysis. Although current approaches attempt to model these interactions via co-attention between histology and genomic data, they focus on only dense local similarity across modalities, which fails to capture global consistency between potential structures, i.e. TME-related interactions of histology and co-expression of genomic data. To address these challenges, we propose a Multimodal Optimal Transport-based Co-Attention Transformer framework with global structure consistency, in which optimal transport (OT) is applied to match patches of a WSI and genes embeddings for selecting informative patches to represent the gigapixel WSI. More importantly, OT-based co-attention provides a global awareness to effectively capture structural interactions within TME for survival prediction. To overcome high computational complexity of OT, we propose a robust and efficient implementation over micro-batch of WSI patches by approximating the original OT with unbalanced mini-batch OT. Extensive experiments show the superiority of our method on five benchmark datasets compared to the state-of-the-art methods. The code will be released.
****
GaFET: Learning Geometry-aware Facial Expression Translation from In-The-Wild Images

Tianxiang Ma, Bingchuan Li, Qian He, Jing Dong, Tieniu Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7115-7125
While current face animation methods can manipulate expressions individually, they suffer from several limitations. The expressions manipulated by some motion-based facial reenactment models are crude. Other ideas modeled with facial action units cannot generalize to arbitrary expressions not covered by annotations. In this paper, we introduce a novel Geometry-aware Facial Expression Translation (GaFET) framework, which is based on parametric 3D facial representations and can stably decoupled expression. Among them, a Multi-level Feature Aligned Transformer is proposed to complement non-geometric facial detail features while addressing the alignment challenge of spatial features. Further, we design a De-expression model based on StyleGAN, in order to reduce the learning difficulty of GaFET in unpaired "in-the-wild" images. Extensive qualitative and quantitative experiments demonstrate that we achieve higher-quality and more accurate facial expression transfer results compared to state-of-the-art methods, and demonstrate applicability of various poses and complex textures. Besides, videos or annotated training data are omitted, making our method easier to use and generalize.
****
Communication-Efficient Vertical Federated Learning with Limited Overlapping Samples
Jingwei Sun, Ziyue Xu, Dong Yang, Vishwesh Nath, Wenqi Li, Can Zhao, Daguang Xu, Yiran Chen, Holger R. Roth; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5203-5212
Federated learning is a popular collaborative learning approach that enables clients to train a global model without sharing their local data. Vertical federated learning (VFL) deals with scenarios in which the data on clients have different feature spaces but share some overlapping samples. Existing VFL approaches suffer from high communication costs and cannot deal efficiently with limited overlapping samples commonly seen in the real world. We propose a practical vertical federated learning (VFL) framework called one-shot VFL that can solve the communication bottleneck and the problem of limited overlapping samples simultaneously based on semi-supervised learning. We also propose few-shot VFL to improve the accuracy further with just one more communication round between the server and the clients. In our proposed framework, the clients only need to communicate with the server once or only a few times. We evaluate the proposed VFL framework on both image and tabular datasets. Our methods can improve the accuracy by more than 46.5% and reduce the communication cost by more than 330xcompared with state-of-the-art VFL methods when evaluated on CIFAR-10. Our code is publicly available.
****
On the Audio-visual Synchronization for Lip-to-Speech Synthesis
Zhe Niu, Brian Mak; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7843-7852
Most lip-to-speech (LTS) synthesis models are trained and evaluated with the assumption that the audio-video pairs in the dataset are well synchronized. In this work, we demonstrate that commonly used audiovisual datasets such as GRID, TCD-TIMIT, and Lip2Wav can, however, have the data asynchrony issue, which will lead to inaccurate evaluation with conventional time alignment-sensitive metrics such as STOI, ESTOI, and MCD. Moreover, training an LTS model with such datasets can result in model asynchrony, meaning that the generated speech and input video are out of sync. To address these problems, we first provide a time-alignment frontend for the commonly used metrics to ensure accurate evaluation. Then, we propose a synchronized lip-to-speech (SLTS) model with an automatic synchronization mechanism (ASM) that corrects data asynchrony and penalizes model asynchrony during training. We evaluated the effectiveness of our approach on both artificial and popular audiovisual datasets. Our proposed method outperforms existing SOTA models in a variety of evaluation metrics.
****
Robust One-Shot Face Video Re-enactment using Hybrid Latent Spaces of StyleGAN2
Trevine Oorloff, Yaser Yacoob; Proceedings of the IEEE/CVF International Confere

nce on Computer Vision (ICCV), 2023, pp. 20947-20957
Recent research on one-shot face re-enactment has progressively overcome the low-resolution constraint with the help of StyleGAN's high-fidelity portrait generation. However, such approaches rely on explicit 2D/3D structural priors for guidance and/or use flow-based warping which constrain their performance. Moreover, existing methods are sensitive (not robust) to the source frame's facial expressions and head pose, even though ideally only the identity of the source frame should have an effect. Addressing these limitations, we propose a novel framework exploiting the implicit 3D prior and inherent latent properties of StyleGAN2 to facilitate one-shot face re-enactment at 1024x1024 (1) with zero dependencies on explicit structural priors, (2) accommodating attribute edits, and (3) robust to diverse facial expressions and head poses of the source frame. We train an encoder using a self-supervised approach to decompose the identity and facial deformation of a portrait image within the pre-trained StyleGAN2's predefined latent spaces itself (automatically facilitating (1) and (2)). The decomposed identity latent of the source and the facial deformation latents of the driving sequence are used to generate re-enacted frames using the StyleGAN2 generator. Additionally, to improve the identity reconstruction and to enable seamless transfer of driving motion, we propose a novel approach, Cyclic Manifold Adjustment. We perform extensive qualitative and quantitative analyses which demonstrate the superiority of the proposed approach against state-of-the-art methods. Project page: https://trevineoorloff.github.io/FaceVideoReenactment_HybridLatents.io/.
****
BallGAN: 3D-aware Image Synthesis with a Spherical Background
Minjung Shin, Yunji Seo, Jeongmin Bae, Young Sun Choi, Hyunsu Kim, Hyeran Byun, Youngjung Uh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7268-7279
3D-aware GANs aim to synthesize realistic 3D scenes that can be rendered in arbitrary camera viewpoints, generating high-quality images with well-defined geometry. As 3D content creation becomes more popular, the ability to generate foreground objects separately from the background has become a crucial property. Existing methods have been developed regarding overall image quality, but they can not generate foreground objects only and often show degraded 3D geometry. In this work, we propose to represent the background as a spherical surface for multiple reasons inspired by computer graphics. Our method naturally provides foreground-only 3D synthesis facilitating easier 3D content creation. Furthermore, it improves the foreground geometry of 3D-aware GANs and the training stability on datasets with complex backgrounds.
****
RPG-Palm: Realistic Pseudo-data Generation for Palmprint Recognition
Lei Shen, Jianlong Jin, Ruixin Zhang, Huaen Li, Kai Zhao, Yingyi Zhang, Jingyun Zhang, Shouhong Ding, Yang Zhao, Wei Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19605-19616
Palmprint recently shows great potential in recognition applications as it is a privacy-friendly and stable biometric. However, the lack of large-scale public palmprint datasets limits further research and development of palmprint recognition. In this paper, we propose a novel realistic pseudo-palmprint generation (RPG) model to synthesize palmprints with massive identities. We first introduce a conditional modulation generator to improve intra-class diversity. Then an identity-aware loss is proposed to ensure identity consistency against unpaired training. We further improve the Bezier palm creases generation strategy to guarantee identity independence. Extensive experimental results demonstrate that synthetic pretraining significantly boosts the recognition model performance. For example, our model improves the state-of-the-art BezierPalm by more than 5% and 14% in terms of TAR@FAR=1e-6 under the 1:1 and 1:3 Open-set protocol. When accessing only 10% of the real training data, our method still outperforms ArcFace with 100% real training data, indicating that we are closer to real-data-free palmprint recognition. The code will be made open upon acceptance.
****
Lecture Presentations Multimodal Dataset: Towards Understanding Multimodality in

Educational Videos
Dong Won Lee, Chaitanya Ahuja, Paul Pu Liang, Sanika Natu, Louis-Philippe Morency; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20087-20098
Many educational videos use slide presentations, a sequence of visual pages that contain text and figures accompanied by spoken language, which are constructed and presented carefully in order to optimally transfer knowledge to students. Previous studies in multimedia and psychology attribute the effectiveness of lecture presentations to their multimodal nature. As a step toward developing vision-language models to aid in student learning as intelligent teacher assistants, we introduce the Lecture Presentations Multimodal (LPM) Dataset as a large-scale benchmark testing the capabilities of vision-and-language models in multimodal understanding of educational videos. Our dataset contains aligned slides and spoken language, for 180+ hours of video and 9000+ slides, with 10 lecturers from various subjects (e.g., computer science, dentistry, biology). We introduce three research tasks, (1) figure-to-text retrieval, (2) text-to-figure retrieval, and (3) generation of slide explanations, which are grounded in multimedia learning and psychology principles to test a vision-language model's understanding of multimodal content. We provide manual annotations to help implement these tasks and establish baselines on them. Comparing baselines and human student performances, we find that state-of-the-art vision-language models (zero-shot and fine-tuned) struggle in (1) weak crossmodal alignment between slides and spoken text, (2) learning novel visual mediums, (3) technical language, and (4) long-range sequences. We introduce PolyViLT, a novel multimodal transformer trained with a multi-instance learning loss that is more effective than current approaches for retrieval. We conclude by shedding light on the challenges and opportunities in multimodal understanding of educational presentation videos.
****
Window-Based Early-Exit Cascades for Uncertainty Estimation: When Deep Ensembles are More Efficient than Single Models
Guoxuan Xia, Christos-Savvas Bouganis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17368-17380
Deep Ensembles are a simple, reliable, and effective method of improving both the predictive performance and uncertainty estimates of deep learning approaches. However, they are widely criticised as being computationally expensive, due to the need to deploy multiple independent models. Recent work has challenged this view, showing that for predictive accuracy, ensembles can be more computationally efficient (at inference) than scaling single models within an architecture family. This is achieved by cascading ensemble members via an early-exit approach. In this work, we investigate extending these efficiency gains to tasks related to uncertainty estimation. As many such tasks, e.g. selective classification, are binary classification, our key novel insight is to only pass samples within a window close to the binary decision boundary to later cascade stages. Experiments on ImageNet-scale data across a number of network architectures and uncertainty tasks show that the proposed window-based early-exit approach is able to achieve a superior uncertainty-computation trade-off compared to scaling single models. For example, a cascaded EfficientNet-B2 ensemble is able to achieve similar coverage at 5% risk as a single EfficientNet-B4 with <30% the number of MACs. We also find that cascades/ensembles give more reliable improvements on OOD data vs scaling models up.
****
AttT2M: Text-Driven Human Motion Generation with Multi-Perspective Attention Mechanism
Chongyang Zhong, Lei Hu, Zihao Zhang, Shihong Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 509-519
Generating 3D human motion based on textual descriptions has been a research focus in recent years. It requires the generated motion to be diverse, natural, and conform to the textual description. Due to the complex spatio-temporal nature of human motion and the difficulty in learning the cross-modal relationship between text and motion, text-driven motion generation is still a challenging problem

. To address these issues, we propose AttT2M, a two-stage method with multi-pers
pective attention mechanism: body-part attention and global-local motion-text at
tention. The former focuses on the motion embedding perspective, which means int
roducing a body-part spatio-temporal encoder into VQ-VAE to learn a more express
ive discrete latent space. The latter is from the cross-modal perspective, which
 is used to learn the sentence-level and word-level motion-text cross-modal rela
tionship. The text-driven motion is finally generated with a generative transfor
mer. Extensive experiments conducted on HumanML3D and KIT-ML demonstrate that ou
r method outperforms the current state-of-the-art works in terms of qualitative
and quantitative evaluation, and achieve fine-grained synthesis and action2motio
n. Our code will be publicly available.
****
A Theory of Topological Derivatives for Inverse Rendering of Geometry
Ishit Mehta, Manmohan Chandraker, Ravi Ramamoorthi; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 419-429
We introduce a theoretical framework for differentiable surface evolution that a
llows discrete topology changes through the use of topological derivatives for v
ariational optimization of image functionals. While prior methods for inverse re
ndering of geometry rely on silhouette gradients for topology changes, such sign
als are sparse. In contrast, our theory derives topological derivatives that rel
ate the introduction of vanishing holes and phases to changes in image intensity
. As a result, we enable differentiable shape perturbations in the form of hole
or phase nucleation. We validate the proposed theory with optimization of closed
 curves in 2D and surfaces in 3D to lend insights into limitations of current me
thods and enable improved applications such as image vectorization, vector-graph
ics generation from text prompts, single-image reconstruction of shape ambigrams
 and multiview 3D reconstruction.
****
Canonical Factors for Hybrid Neural Fields
Brent Yi, Weijia Zeng, Sam Buchanan, Yi Ma; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 3414-3426
Factored feature volumes offer a simple way to build more compact, efficient, an
d intepretable neural fields, but also introduce biases that are not necessarily
 beneficial for real-world data. In this work, we (1) characterize the undesirab
le biases that these architectures have for axis-aligned signals -- they can lea
d to radiance field reconstruction differences of as high as 2 PSNR -- and (2) e
xplore how learning a set of canonicalizing transformations can improve represen
tations by removing these biases. We prove in a simple two-dimensional model pro
blem that a hybrid architecture that simultaneously learns these transformations
 together with scene appearance succeeds with drastically improved efficiency. W
e validate the resulting architectures, which we call TILTED, using 2D image, si
gned distance field, and radiance field reconstruction tasks, where we observe i
mprovements across quality, robustness, compactness, and runtime. Results demons
trate that TILTED can enable capabilities comparable to baselines that are 2x la
rger, while highlighting weaknesses of standard procedures for evaluating neural
 field representations.
****
XNet: Wavelet-Based Low and High Frequency Fusion Networks for Fully- and Semi-S
upervised Semantic Segmentation of Biomedical Images
Yanfeng Zhou, Jiaxing Huang, Chenlong Wang, Le Song, Ge Yang; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21085-21
096
Fully- and semi-supervised semantic segmentation of biomedical images have been
advanced with the development of deep neural networks (DNNs). So far, however, D
NN models are usually designed to support one of these two learning schemes, uni
fied models that support both fully- and semi-supervised segmentation remain lim
ited. Furthermore, few fully-supervised models focus on the intrinsic low freque
ncy (LF) and high frequency (HF) information of images to improve performance. P
erturbations in consistency-based semi-supervised models are often artificially
designed. They may introduce negative learning bias that are not beneficial for

training. In this study, we propose a wavelet-based LF and HF fusion model XNet, which supports both fully- and semi-supervised semantic segmentation and outperforms state-of-the-art models in both fields. It emphasizes extracting LF and HF information for consistency training to alleviate the learning bias caused by artificial perturbations. Extensive experiments on two 2D and two 3D datasets demonstrate the effectiveness of our model. Code is available at https://github.com/Yanfeng-Zhou/XNet.
****
Betrayed by Captions: Joint Caption Grounding and Generation for Open Vocabulary Instance Segmentation
Jianzong Wu, Xiangtai Li, Henghui Ding, Xia Li, Guangliang Cheng, Yunhai Tong, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21938-21948
In this work, we focus on open vocabulary instance segmentation to expand a segmentation model to classify and segment instance-level novel categories. Previous approaches have relied on massive caption datasets and complex pipelines to establish one-to-one mappings between image regions and words in captions. However, such methods build noisy supervision by matching non-visible words to image regions, such as adjectives and verbs. Meanwhile, context words are also important for inferring the existence of novel objects as they show high inter-correlations with novel categories. To overcome these limitations, we devise a joint Caption Grounding and Generation (CGG) framework, which incorporates a novel grounding loss that only focuses on matching object nouns to improve learning efficiency. We also introduce a caption generation head that enables additional supervision and contextual modeling as a complementation to the grounding loss. Our analysis and results demonstrate that grounding and generation components complement each other, significantly enhancing the segmentation performance for novel classes. Experiments on the COCO dataset with two settings: Open Vocabulary Instance Segmentation (OVIS) and Open Set Panoptic Segmentation (OSPS) demonstrate the superiority of the CGG. Specifically, CGG achieves a substantial improvement of 6.8% mAP for novel classes without extra data on the OVIS task and 15% PQ improvements for novel classes on the OSPS benchmark.
****
StyleGANEX: StyleGAN-Based Manipulation Beyond Cropped Aligned Faces
Shuai Yang, Liming Jiang, Ziwei Liu, Chen Change Loy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21000-21010
Recent advances in face manipulation using StyleGAN have produced impressive results. However, StyleGAN is inherently limited to cropped aligned faces at a fixed image resolution it is pre-trained on. In this paper, we propose a simple and effective solution to this limitation by using dilated convolutions to rescale the receptive fields of shallow layers in StyleGAN, without altering any model parameters. This allows fixed-size small features at shallow layers to be extended into larger ones that can accommodate variable resolutions, making them more robust in characterizing unaligned faces. To enable real face inversion and manipulation, we introduce a corresponding encoder that provides the first-layer feature of the extended StyleGAN in addition to the latent style code. We validate the effectiveness of our method using unaligned face inputs of various resolutions in a diverse set of face manipulation tasks, including facial attribute editing, super-resolution, sketch/mask-to-face translation, and face toonification.
****
HandR2N2: Iterative 3D Hand Pose Estimation Using a Residual Recurrent Neural Network
Wencan Cheng, Jong Hwan Ko; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20904-20913
3D hand pose estimation is a critical task in various human-computer interaction applications. Numerous deep learning based estimation models in this domain have been actively explored. However, the existing models follows a non-recurrent scheme and thus require complex architectures or redundant parameters in order to achieve acceptable model capacity. To tackle this limitation, this paper proposes HandR2N2, a compact neural network that iteratively regresses the hand pose u

sing a novel residual recurrent unit. The recurrent design allows recursive expl
oitation of partial layers to gradually optimize previously estimated joint loca
tions. In addition, we exploit graph reasoning to capture kinematic dependencies
 between joints for better performance. Experimental results show that the propo
sed model significantly outperforms the existing methods on three hand pose benc
hmark datasets in terms of both accuracy and efficiency. Codes and pre-trained m
odels are publicly available at https://github.com/cwc1260/HandR2N2.
****
GET: Group Event Transformer for Event-Based Vision
Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, Feng Wu; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6038-60
48
Event cameras are a type of novel neuromorphic sen-sor that has been gaining inc
reasing attention. Existing event-based backbones mainly rely on image-based des
igns to extract spatial information within the image transformed from events, ov
erlooking important event properties like time and polarity. To address this iss
ue, we propose a novel Group-based vision Transformer backbone for Event-based v
ision, called Group Event Transformer (GET), which de-couples temporal-polarity
information from spatial infor-mation throughout the feature extraction process.
 Specifi-cally, we first propose a new event representation for GET, named Group
 Token, which groups asynchronous events based on their timestamps and polaritie
s. Then, GET ap-plies the Event Dual Self-Attention block, and Group Token Aggre
gation module to facilitate effective feature commu-nication and integration in
both the spatial and temporal-polarity domains. After that, GET can be integrate
d with different downstream tasks by connecting it with vari-ous heads. We evalu
ate our method on four event-based classification datasets (Cifar10-DVS, N-MNIST
, N-CARS, and DVS128Gesture) and two event-based object detection datasets (1Mpx
 and Gen1), and the results demonstrate that GET outperforms other state-of-the-
art methods. The code is available at https://github.com/Peterande/GET-Group-Eve
nt-Transformer.
****
Unsupervised Learning of Object-Centric Embeddings for Cell Instance Segmentatio
n in Microscopy Images
Steffen Wolf, Manan Lalit, Katie McDole, Jan Funke; Proceedings of the IEEE/CVF
International Conference on Computer Vision (ICCV), 2023, pp. 21263-21272
Segmentation of objects in microscopy images is required for many biomedical app
lications. We introduce object-centric embeddings (OCEs), which embed image patc
hes such that the spatial offsets between patches cropped from the same object a
re preserved. Those learnt embeddings can be used to delineate individual object
s and thus obtain instance segmentations. Here, we show theoretically that, unde
r assumptions commonly found in microscopy images, OCEs can be learnt through a
self-supervised task that predicts the spatial offset between image patches. Tog
ether, this forms an unsupervised cell instance segmentation method which we eva
luate on nine diverse large-scale microscopy datasets. Segmentations obtained wi
th our method lead to substantially improved results, compared to a state-of-the
-art baseline on six out of nine datasets, and perform on par on the remaining t
hree datasets. If ground-truth annotations are available, our method serves as a
n excellent starting point for supervised training, reducing the required amount
 of ground-truth needed by one order of magnitude, thus substantially increasing
 the practical applicability of our method. Source code is available at github.c
om/funkelab/cellulus.
****
DyGait: Exploiting Dynamic Representations for High-performance Gait Recognition
Ming Wang, Xianda Guo, Beibei Lin, Tian Yang, Zheng Zhu, Lincheng Li, Shunli Zha
ng, Xin Yu; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2023, pp. 13424-13433
Gait recognition is a biometric technology that recognizes the identity of human
s through their walking patterns. Compared with other biometric technologies, ga
it recognition is more difficult to disguise and can be applied to the condition
 of long-distance without the cooperation of subjects. Thus, it has unique poten

tial and wide application for crime prevention and social security. At present, most gait recognition methods directly extract features from the video frames to establish representations. However, these architectures learn representations from different features equally but do not pay enough attention to dynamic features, which refers to a representation of dynamic parts of silhouettes over time (e.g. legs). Since dynamic parts of the human body are more informative than other parts (e.g. bags) during walking, in this paper, we propose a novel and high-performance framework named DyGait. This is the first framework on gait recognition that is designed to focus on the extraction of dynamic features. Specifically, to take full advantage of the dynamic information, we propose a Dynamic Augmentation Module (DAM), which can automatically establish spatial-temporal feature representations of the dynamic parts of the human body. The experimental results show that our DyGait network outperforms other state-of-the-art gait recognition methods. It achieves an average Rank-1 accuracy of 71.4% on the GREW dataset, 66.3% on the Gait3D dataset, 98.4% on the CASIA-B dataset and 98.3% on the OU-MVLP dataset.
****
When Do Curricula Work in Federated Learning?
Saeed Vahidian, Sreevatsank Kadaveru, Woonjoon Baek, Weijia Wang, Vyacheslav Kungurtsev, Chen Chen, Mubarak Shah, Bill Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5084-5094
An oft-cited open problem of federated learning is the existence of data heterogeneity among clients. One path- way to understanding the drastic accuracy drop in feder- ated learning is by scrutinizing the behavior of the clients' deep models on data with different levels of "difficulty", which has been left unaddressed. In this paper, we investi- gate a different and rarely studied dimension of FL: ordered learning. Specifically, we aim to investigate how ordered learning principles can contribute to alleviating the hetero- geneity effects in FL. We present theoretical analysis and conduct extensive empirical studies on the efficacy of or- derings spanning three kinds of learning: curriculum, anti- curriculum, and random curriculum. We find that curricu- lum learning largely alleviates non-IIDness. Interestingly, the more disparate the data distributions across clients the more they benefit from ordered learning. We provide analysis explaining this phenomenon, specifically indicating how curriculum training appears to make the objective land- scape progressively less convex, suggesting fast converging iterations at the beginning of the training procedure. We derive quantitative results of convergence for both convex and nonconvex objectives by modeling the curriculum train- ing on federated devices as local SGD with locally biased stochastic gradients. Also, inspired by ordered learning, we propose a novel client selection technique that benefits from the real-world disparity in the clients. Our proposed approach to client selection has a synergic effect when applied together with ordered learning in FL.
****
XiNet: Efficient Neural Networks for tinyML
Alberto Ancilotto, Francesco Paissan, Elisabetta Farella; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16968-16977
The recent interest in the edge-to-cloud continuum paradigm has emphasized the need for simple and scalable architectures to deliver optimal performance on computationally constrained devices. However, resource-efficient neural networks usually optimize for parameter count and thus use operators such as depthwise convolutions, which do not maximally exploit the efficiency of resource-constrained devices. In this article, we propose XiNet, a novel convolutional neural architecture that targets edge devices. We derived the XiNet architecture from an extensive real-world efficiency analysis of various neural network operators (e.g., standard, depthwise, and pointwise convolutions). Compared to other mobile architectures, our approach substantially improves the performance-complexity trade-off by optimizing the number of operations, parameters, and working memory (RAM). Moreover, we show how XiNet can be easily adapted to different devices thanks to Hardware Aware Scaling (HAS), which enables disjoint optimization of RAM, FLASH, and operations count. We analyze the scaling properties of our architecture und

er different hardware constraints and validate it on the image classification task. Finally, we evaluate the performance of XiNet for object detection on the MS-COCO and VOC-2012 benchmarks and compare it with state-of-the-art mobile neural networks, achieving a 70% reduction in energy requirements with similar performance.
****
GridPull: Towards Scalability in Learning Implicit Representations from 3D Point Clouds
Chao Chen, Yu-Shen Liu, Zhizhong Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18322-18334
Learning implicit representations has been a widely used solution for surface reconstruction from 3D point clouds. The latest methods infer a distance or occupancy field by overfitting a neural network on a single point cloud. However, these methods suffer from a slow inference due to the slow convergence of neural networks and the extensive calculation of distances to surface points, which limits them to small scale points. To resolve the scalability issue in surface reconstruction, we propose GridPull to improve the efficiency of learning implicit representations from large scale point clouds. Our novelty lies in the fast inference of a discrete distance field defined on grids without using any neural components. To remedy the lack of continuousness brought by neural networks, we introduce a loss function to encourage continuous distances and consistent gradients in the field during pulling queries onto the surface in grids near to the surface. We use uniform grids for a fast grid search to localize sampled queries, and organize surface points in a tree structure to speed up the calculation of distances to the surface. We do not rely on learning priors or normal supervision during optimization, and achieve superiority over the latest methods in terms of complexity and accuracy. We evaluate our method on shape and scene benchmarks, and report numerical and visual comparisons with the latest methods to justify our effectiveness and superiority. The code is available at https://github.com/chenchao15/GridPull.
****
Audio-Visual Class-Incremental Learning
Weiguo Pian, Shentong Mo, Yunhui Guo, Yapeng Tian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7799-7811
In this paper, we introduce audio-visual class-incremental learning, a class-incremental learning scenario for audio-visual video recognition. We demonstrate that joint audio-visual modeling can improve class-incremental learning, but current methods fail to preserve semantic similarity between audio and visual features as incremental step grows. Furthermore, we observe that audio-visual correlations learned in previous tasks can be forgotten as incremental steps progress, leading to poor performance. To overcome these challenges, we propose AV-CIL, which incorporates Dual-Audio-Visual Similarity Constraint (D-AVSC) to maintain both instance-aware and class-aware semantic similarity between audio-visual modalities and Visual Attention Distillation (VAD) to retain previously learned audio-guided visual attentive ability. We create three audio-visual class-incremental datasets, AVE-Class-Incremental (AVE-CI), Kinetics-Sounds-Class-Incremental (K-S-CI), and VGGSound100-Class-Incremental (VS100-CI) based on the AVE, Kinetics-Sounds, and VGGSound datasets, respectively. Our experiments on AVE-CI, K-S-CI, and VS100-CI demonstrate that AV-CIL significantly outperforms existing class-incremental learning methods in audio-visual class-incremental learning. Code and data are available at: https://github.com/weiguoPian/AV-CIL_ICCV2023.
****
GeoMIM: Towards Better 3D Knowledge Transfer via Masked Image Modeling for Multi-view 3D Understanding
Jihao Liu, Tai Wang, Boxiao Liu, Qihang Zhang, Yu Liu, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17839-17849
Multi-view camera-based 3D detection is a challenging problem in computer vision. Recent works leverage a pretrained LiDAR detection model to transfer knowledge to a camera-based student network. However, we argue that there is a major doma

in gap between the LiDAR BEV features and the camera-based BEV features, as they have different characteristics and are derived from different sources. In this paper, we propose Geometry Enhanced Masked Image Modeling (GeoMIM) to transfer the knowledge of the LiDAR model in a pretrain-finetune paradigm for improving the multi-view camera-based 3D detection. GeoMIM is a multi-camera vision transformer with Cross-View Attention (CVA) blocks that uses LiDAR BEV features encoded by the pretrained BEV model as learning targets. During pretraining, GeoMIM's decoder has a semantic branch completing dense perspective-view features and the other geometry branch reconstructing dense perspective-view depth maps. The depth branch is designed to be camera-aware by inputting the camera's parameters for better transfer capability. Extensive results demonstrate that GeoMIM outperforms existing methods on nuScenes benchmark, achieving state-of-the-art performance for camera-based 3D object detection and 3D segmentation.
****
Towards Viewpoint-Invariant Visual Recognition via Adversarial Training
Shouwei Ruan, Yinpeng Dong, Hang Su, Jianteng Peng, Ning Chen, Xingxing Wei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4709-4719
Visual recognition models are not invariant to viewpoint changes in the 3D world, as different viewing directions can dramatically affect the predictions given the same object. Although many efforts have been devoted to making neural networks invariant to 2D image translations and rotations, viewpoint invariance is rarely investigated. As most models process images in the perspective view, it is challenging to impose invariance to 3D viewpoint changes based only on 2D inputs. Motivated by the success of adversarial training in promoting model robustness, we propose Viewpoint-Invariant Adversarial Training (VIAT) to improve viewpoint robustness of common image classifiers. By regarding viewpoint transformation as an attack, VIAT is formulated as a minimax optimization problem, where the inner maximization characterizes diverse adversarial viewpoints by learning a Gaussian mixture distribution based on a new attack GMVFool, while the outer minimization trains a viewpoint-invariant classifier by minimizing the expected loss over the worst-case adversarial viewpoint distributions. To further improve the generalization performance, a distribution sharing strategy is introduced leveraging the transferability of adversarial viewpoints across objects. Experiments validate the effectiveness of VIAT in improving the viewpoint robustness of various image classifiers based on the diversity of adversarial viewpoints generated by GMVFool.
****
Helping Hands: An Object-Aware Ego-Centric Video Recognition Model
Chuhan Zhang, Ankush Gupta, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13901-13912
We introduce an object-aware decoder for improving the performance of spatio-temporal representations on ego-centric videos. The key idea is to enhance object-awareness during training by tasking the model to predict hand positions, object positions, and the semantic label of the objects using paired captions when available. At inference time the model only requires RGB frames as inputs, and is able to track and ground objects (although it has not been trained explicitly for this).

We demonstrate the performance of the object-aware representations learnt by our model, by: (i) evaluating it for strong transfer, i.e, through zero-shot testing, on a number of downstream video-text retrieval and classification benchmarks; and (ii) by evaluating its temporal and spatial (grounding) performance by fine-tuning for this task.

In all cases the performance improves over the state of the art -- even for networks trained with far larger batch sizes. Overall, we show that the model can act as a drop-in replacement for an ego-centric video model, and improve performance.
****

RenderIH: A Large-Scale Synthetic Dataset for 3D Interacting Hand Pose Estimation

Lijun Li, Linrui Tian, Xindi Zhang, Qi Wang, Bang Zhang, Liefeng Bo, Mengyuan Liu, Chen Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20395-20405

The current interacting hand (IH) datasets are relatively simplistic in terms of background and texture, with hand joints being annotated by a machine annotator, which may result in inaccuracies, and the diversity of pose distribution is limited. However, the variability of background, pose distribution, and texture can greatly influence the generalization ability. Therefore, we present a large-scale synthetic dataset --RenderIH-- for interacting hands with accurate and diverse pose annotations. The dataset contains 1M photo-realistic images with varied backgrounds, perspectives, and hand textures. To generate natural and diverse interacting poses, we propose a new pose optimization algorithm. Additionally, for better pose estimation accuracy, we introduce a transformer-based pose estimation network, TransHand, to leverage the correlation between interacting hands and verify the effectiveness of RenderIH in improving results. Our dataset is model-agnostic and can improve more accuracy of any hand pose estimation method in comparison to other real or synthetic datasets. Experiments have shown that pretraining on our synthetic data can significantly decrease the error from 6.76mm to 5.79mm, and our Transhand surpasses contemporary methods. Our dataset and code are available at https://github.com/adwardlee/RenderIH.

****

Multi-Metrics Adaptively Identifies Backdoors in Federated Learning

Siquan Huang, Yijiang Li, Chong Chen, Leyu Shi, Ying Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4652-4662

The decentralized and privacy-preserving nature of federated learning (FL) makes it vulnerable to backdoor attacks aiming to manipulate the behavior of the resulting model on specific adversary-chosen inputs. However, most existing defenses based on statistical differences take effect only against specific attacks, especially when the malicious gradients are similar to benign ones or the data are highly non-independent and identically distributed (non-IID). In this paper, we revisit the distance-based defense methods and discover that i) Euclidean distance becomes meaningless in high dimensions and ii) malicious gradients with diverse characteristics cannot be identified by a single metric. To this end, we present a simple yet effective defense strategy with multi-metrics and dynamic weighting to identify backdoors adaptively. Furthermore, our novel defense has no reliance on predefined assumptions over attack settings or data distributions and little impact on benign performance. To evaluate the effectiveness of our approach, we conduct comprehensive experiments on different datasets under various attack settings, where our method achieves the best defensive performance. For instance, we achieve the lowest backdoor accuracy of 3.06% under the difficult Edge-case PGD, showing significant superiority over previous defenses. The results also demonstrate that our method can be well-adapted to a wide range of non-IID degrees without sacrificing the benign performance.

****

SpinCam: High-Speed Imaging via a Rotating Point-Spread Function

Dorian Chan, Mark Sheinin, Matthew O'Toole; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10789-10799

High-speed cameras are an indispensable tool used for the slow-motion analysis of scenes. The fixed bandwidth of any imaging system quickly becomes a bottleneck however, resulting in a fundamental trade-off between the camera's spatial and temporal resolutions. In recent years, compressive high-speed imaging systems have been proposed to circumvent these issues, by optically compressing the signal and using a reconstruction procedure to recover a video. In our work, we propose a novel approach for compressive high-speed imaging based on temporally coding the camera's point-spread function (PSF). By mechanically spinning a diffraction grating in front of a camera, the sensor integrates an image blurred by a PSF that continuously rotates over time. We also propose a deconvolution-based reconstruction algorithm to reconstruct videos from these measurements. Our method ac

hieves superior light efficiency and handles a wider class of scenes compared to prior methods. Also, our mechanical design yields flexible temporal resolution that can be easily increased, potentially allowing capture at 192 kHz--far higher than prior works. We demonstrate a prototype on various applications including motion capture and particle image velocimetry (PIV).
****
FPR: False Positive Rectification for Weakly Supervised Semantic Segmentation
Liyi Chen, Chenyang Lei, Ruihuang Li, Shuai Li, Zhaoxiang Zhang, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1108-1118
Many weakly supervised semantic segmentation (WSSS) methods employ the class activation map (CAM) to generate the initial segmentation results. However, CAM often fails to distinguish the foreground from its co-occurred background (e.g., train and railroad), resulting in inaccurate activation from the background. Previous endeavors address this co-occurrence issue by introducing external supervision and human priors. In this paper, we present a False Positive Rectification (FPR) approach to tackle the co-occurrence problem by leveraging the false positives of CAM. Based on the observation that the CAM-activated regions of absent classes contain class-specific co-occurred background cues, we collect these false positives and utilize them to guide the training of CAM network by proposing a region-level contrast loss and a pixel-level rectification loss. Without introducing any external supervision and human priors, the proposed FPR effectively suppresses wrong activations from the background objects. Extensive experiments on the PASCAL VOC 2012 and MS COCO 2014 demonstrate that FPR brings significant improvements for off-the-shelf methods and achieves state-of-the-art performance. Code is available at https://github.com/mt-cly/FPR.
****
Cross-modal Scalable Hierarchical Clustering in Hyperbolic space
Teng Long, Nanne van Noord; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16655-16664
Hierarchical clustering is a natural approach to discover ontologies from data. Yet, existing approaches are hampered by their inability to scale to large datasets and the discrete encoding of the hierarchy. We introduce scalable Hyperbolic Hierarchical Clustering (sHHC) which overcomes these limitations by learning continuous hierarchies in hyperbolic space. Our hierarchical clustering is of high quality and can be obtained in a fraction of the runtime.

Additionally, we demonstrate the strength of sHHC on a downstream cross-modal self-supervision task. By using the discovered hierarchies from sound and vision to construct continuous hierarchical pseudo-labels we can efficiently optimize a network for activity recognition and obtain competitive performance compared to recent self-supervised learning models. Our findings demonstrate the strength of Hyperbolic Hierarchical Clustering and its potential for Self-Supervised Learning.
****
DETRDistill: A Universal Knowledge Distillation Framework for DETR-families
Jiahao Chang, Shuo Wang, Hai-Ming Xu, Zehui Chen, Chenhongyi Yang, Feng Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6898-6908
Transformer-based detectors (DETRs) are becoming popular for their simple framework, but the large model size and heavy time consumption hinder their deployment in the real world. While knowledge distillation (KD) can be an appealing technique to compress giant detectors into small ones for comparable detection performance and low inference cost. Since DETRs formulate object detection as a set prediction problem, existing KD methods designed for classic convolution-based detectors may not be directly applicable. In this paper, we propose DETRDistill, a novel knowledge distillation method dedicated to DETR-families. Specifically, we first design a Hungarian-matching logits distillation to encourage the student model to have the exact predictions as those of the teacher DETRs. Then, we propose a target-aware feature distillation to help the student model learn from the

object-centric features of the teacher model. Finally, in order to improve the c
onvergence rate of the student DETR, we introduce a query-prior assignment disti
llation to speed up the student model learning from well-trained queries and sta
ble assignment of the teacher model. Extensive experimental results on the COCO
dataset validate the effectiveness of our approach. Notably, DETRDistill consist
ently improves various DETRs by more than 2.0 mAP, even surpassing their teacher
 models.
****
F&F Attack: Adversarial Attack against Multiple Object Trackers by Inducing Fals
e Negatives and False Positives
Tao Zhou, Qi Ye, Wenhan Luo, Kaihao Zhang, Zhiguo Shi, Jiming Chen; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 45
73-4583
Multi-object tracking (MOT) aims to build moving trajectories for number-agnosti
c objects. Modern multi-object trackers commonly follow the tracking-by-detectio
n strategy. Therefore, fooling detectors can be an effective solution but it usu
ally requires attacks in multiple successive frames, resulting in low efficiency
. Attacking association processes improves efficiency but may require model-spec
ific design, leading to poor generalization. In this paper, we propose a novel F
alse negative and False positive attack (F&F attack) mechanism: it perturbs the
input image to erase original detections and to inject deceptive false alarms ar
ound original ones while integrating the association attack implicitly. The mech
anism can produce effective identity switches against multi-object trackers by o
nly fooling detectors in a few frames. To demonstrate the flexibility of the mec
hanism, we deploy it to three multi-object trackers (ByteTrack, SORT, and Center
Track) which are enabled by two representative detectors (YOLOX and CenterNet).
Comprehensive experiments on MOT17 and MOT20 datasets show that our method signi
ficantly outperforms existing attackers, revealing the vulnerability of the trac
king-by-detection paradigm to detection attacks.
****
Transferable Decoding with Visual Entities for Zero-Shot Image Captioning
Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, Feng Zheng; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
 pp. 3136-3146
Image-to-text generation aims to describe images using natural language. Recentl
y, zero-shot image captioning based on pre-trained vision-language models (VLMs)
 and large language models (LLMs) has made significant progress. However, we hav
e observed and empirically demonstrated that these methods are susceptible to mo
dality bias induced by LLMs and tend to generate descriptions containing objects
 (entities) that do not actually exist in the image but frequently appear during
 training (i.e., object hallucination). In this paper, we propose ViECap, a tran
sferable decoding model that leverages entity-aware decoding to generate descrip
tions in both seen and unseen scenarios. ViECap incorporates entity-aware hard p
rompts to guide LLMs' attention toward the visual entities present in the image,
 enabling coherent caption generation across diverse scenes. With entity-aware h
ard prompts, ViECap is capable of maintaining performance when transferring from
 in-domain to out-of-domain scenarios. Extensive experiments demonstrate that Vi
ECap sets a new state-of-theart cross-domain (transferable) captioning and perfo
rms competitively in-domain captioning compared to previous VLMs-based zero-shot
 methods. Our code is available at: https://github.com/FeiElysia/ViECap
****
ReMoDiffuse: Retrieval-Augmented Motion Diffusion Model
Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li,
 Lei Yang, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Co
mputer Vision (ICCV), 2023, pp. 364-373
3D human motion generation is crucial for creative industry. Recent advances rel
y on generative models with domain knowledge for text-driven motion generation,
leading to substantial progress in capturing common motions. However, the perfor
mance on more diverse motions remains unsatisfactory. In this work, we propose R
eMoDiffuse, a diffusion-model-based motion generation framework that integrates

a retrieval mechanism to refine the denoising process. ReMoDiffuse enhances the generalizability and diversity of text-driven motion generation with three key designs:1) Hybrid Retrieval finds appropriate references from the database in terms of both semantic and kinematic similarities. 2) Semantic-Modulated Transformer selectively absorbs retrieval knowledge, adapting to the difference between retrieved samples and the target motion sequence. 3) Condition Mixture better utilizes the retrieval database during inference, overcoming the scale sensitivity in classifier-free guidance. Extensive experiments demonstrate that \name outperforms state-of-the-art methods by balancing both text-motion consistency and motion quality, especially for more diverse motion generation.
****
GlueStick: Robust Image Matching by Sticking Points and Lines Together
Rémi Pautrat, Iago Suárez, Yifan Yu, Marc Pollefeys, Viktor Larsson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9706-9716
Line segments are powerful features complementary to points. They offer structural cues, robust to drastic viewpoint and illumination changes, and can be present even in texture-less areas. However, describing and matching them is more challenging compared to points due to partial occlusions, lack of texture, or repetitiveness. This paper introduces a new matching paradigm, where points, lines, and their descriptors are unified into a single wireframe structure. We propose GlueStick, a deep matching Graph Neural Network (GNN) that takes two wireframes from different images and leverages the connectivity information between nodes to better glue them together. In addition to the increased efficiency brought by the joint matching, we also demonstrate a large boost of performance when leveraging the complementary nature of these two features in a single architecture. We show that our matching strategy outperforms the state-of-the-art approaches independently matching line segments and points for a wide variety of datasets and tasks. Code is available at https://github.com/cvg/GlueStick.
****
Computational 3D Imaging with Position Sensors
Jeremy Klotz, Mohit Gupta, Aswin C. Sankaranarayanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8125-8134
Underlying many structured light systems, especially those based on laser scanning, is a simple vision task: tracking a light spot. To accomplish this, scanners use conventional CMOS sensors to capture, transmit, and process millions of pixel measurements. This approach, while capable of achieving high-fidelity 3D scans, is wasteful in terms of (often scarce) sensing and computational resources. We present a structured light system based on position sensing diodes (PSDs), an unconventional sensing modality that directly measures the centroid of the spatial distribution of incident light, thus enabling high-resolution 3D laser scanning with a minimal amount of sensor data. We develop theory and computational algorithms for PSD-based structured light under a variety of light transport effects. We demonstrate the benefits of the proposed techniques using a hardware prototype on several real-world scenes, including optically-challenging objects with long-range inter-reflections and scattering.
****
PointMBF: A Multi-scale Bidirectional Fusion Network for Unsupervised RGB-D Point Cloud Registration
Mingzhi Yuan, Kexue Fu, Zhihao Li, Yucong Meng, Manning Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17694-17705
Point cloud registration is a task to estimate the rigid transformation between two unaligned scans, which plays an important role in many computer vision applications. Previous learning-based works commonly focus on supervised registration, which have limitations in practice. Recently, with the advance of inexpensive RGB-D sensors, several learning-based works utilize RGB-D data to achieve unsupervised registration. However, most of existing unsupervised methods follow a cascaded design or fuse RGB-D data in a unidirectional manner, which do not fully exploit the complementary information in the RGB-D data. To leverage the compleme

ntary information more effectively, we propose a network implementing multi-scale bidirectional fusion between RGB images and point clouds generated from depth images. By bidirectionally fusing visual and geometric features in multi-scales, more distinctive deep features for correspondence estimation can be obtained, making our registration more accurate. Extensive experiments on ScanNet and 3DMatch demonstrate that our method achieves new state-of-the-art performance. Code will be released at https://github.com/phdymz/PointMBF.
****
Towards Multi-Layered 3D Garments Animation
Yidi Shao, Chen Change Loy, Bo Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14361-14370
Mimicking realistic dynamics in 3D garment animations is a challenging task due to the complex nature of multi-layered garments and the variety of outer forces involved. Existing approaches mostly focus on single-layered garments driven by only human bodies and struggle to handle general scenarios. In this paper, we propose a novel data-driven method, called LayersNet, to model garment-level animations as particle-wise interactions in a micro physics system. We improve simulation efficiency by representing garments as patch-level particles in a two-level structural hierarchy. Moreover, we introduce a novel Rotation Equivalent Transformation with Rotation Invariant Attention that leverage the rotation invariance and additivity of physics systems to better model outer forces. To verify the effectiveness of our approach and bridge the gap between experimental environments and real-world scenarios, we introduce a new challenging dataset, D-LAYERS, containing 700K frames of dynamics of 4,900 combinations of multi-layered garments driven by human bodies and randomly sampled wind. Our LayersNet achieves superior performance both quantitatively and qualitatively. Project page: www.mmlab-ntu.com/project/layersnet/index.html .
****
LiveHand: Real-time and Photorealistic Neural Hand Rendering
Akshay Mundra, Mallikarjun B R, Jiayi Wang, Marc Habermann, Christian Theobalt, Mohamed Elgharib; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18035-18045
The human hand is the main medium through which we interact with our surroundings, making its digitization an important problem. While there are several works modeling the geometry of hands, little attention has been paid to capturing photo-realistic appearance. Moreover, for applications in extended reality and gaming, real-time rendering is critical. We present the first neural-implicit approach to photo-realistically render hands in real-time. This is a challenging problem as hands are textured and undergo strong articulations with pose-dependent effects. However, we show that this aim is achievable through our carefully designed method. This includes training on a low-resolution rendering of a neural radiance field, together with a 3D-consistent super-resolution module and mesh-guided sampling and space canonicalization. We demonstrate a novel application of perceptual loss on the image space, which is critical for learning details accurately. We also show a live demo where we photo-realistically render the human hand in real-time for the first time, while also modeling pose- and view-dependent appearance effects. We ablate all our design choices and show that they optimize for rendering speed and quality.
****
Advancing Referring Expression Segmentation Beyond Single Image
Yixuan Wu, Zhao Zhang, Chi Xie, Feng Zhu, Rui Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2628-2638
Referring Expression Segmentation (RES) is a widely explored multi-modal task, which endeavors to segment the pre-existing object within a single image with a given linguistic expression. However, in broader real-world scenarios, it is not always possible to determine if the described object exists in a specific image. Generally, a collection of images is available, some of which potentially contain the target objects. To this end, we propose a more realistic setting, named Group-wise Referring Expression Segmentation (GRES), which expands RES to a group of related images, allowing the described objects to exist in a subset of the i

nput image group. To support this new setting, we introduce an elaborately compiled dataset named Grouped Referring Dataset (GRD), containing complete group-wise annotations of the target objects described by given expressions. Moreover, we also present a baseline method named Grouped Referring Segmenter (GRSer), which explicitly captures the language-vision and intra-group vision-vision interactions to achieve state-of-the-art results on the proposed GRES setting and related tasks, such as Co-Salient Object Detection and traditional RES. Our dataset and codes are publicly released in https://github.com/shikras/d-cube.
****
Learning Image Harmonization in the Linear Color Space
Ke Xu, Gerhard Petrus Hancke, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12570-12579
Harmonizing cut-and-paste images into perceptually realistic ones is challenging, as it requires a full understanding of the discrepancies between the background of the target image and the inserted object. Existing methods mainly adjust the appearances of the inserted object via pixel-level manipulations. They are not effective in correcting color discrepancy caused by different scene illuminations and the image formation processes. We note that image colors are essentially camera ISP projection of the scene radiance. If we can trace the image colors back to the radiance field, we may be able to model the scene illumination and harmonize the discrepancy better. In this paper, we propose a novel neural approach to harmonize the image colors in a camera-independent color space, in which color values are proportional to the scene radiance. To this end, we propose a novel image unprocessing module to estimate an intermediate high dynamic range version of the object to be inserted. We then propose a novel color harmonization module that harmonizes the colors of the inserted object by querying the estimated scene radiance and re-rendering the harmonized object in the output color space. Extensive experiments demonstrate that our method outperforms the state-of-the-art approaches.
****
Chasing Clouds: Differentiable Volumetric Rasterisation of Point Clouds as a Highly Efficient and Accurate Loss for Large-Scale Deformable 3D Registration
Mattias P. Heinrich, Alexander Bigalke, Christoph Großbröhmer, Lasse Hansen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8026-8036
Learning-based registration for large-scale 3D point clouds has been shown to improve robustness and accuracy compared to classical methods and can be trained without supervision for locally rigid problems. However, for tasks with highly deformable structures, such as alignment of pulmonary vascular trees for medical diagnostics, previous approaches of self-supervision with regularisation and point distance losses have failed to succeed, leading to the need for complex synthetic augmentation strategies to obtain reliably strong supervision. In this work, we introduce a novel Differentiable Volumetric Rasterisation of point Clouds (DiVRoC) that overcomes those limitations and offers a highly efficient and accurate loss for large-scale deformable 3D registration. DiVRoC drastically reduces the computational complexity for measuring point cloud distances for high-resolution data with over 100k 3D points and can also be employed to extrapolate and regularise sparse motion fields, as loss in a self-training setting and as objective function in instance optimisation. DiVRoC can be successfully embedded into geometric registration networks, including PointPWC-Net and other graph CNNs. Our approach yields new state-of-the-art accuracy on the challenging PVT dataset in three different settings without training with manual ground truth: 1) unsupervised metric-based learning 2) self-supervised learning with pseudo labels generated by self-training and 3) optimisation based alignment without learning. https://github.com/mattiaspaul/ChasingClouds
****
TripLe: Revisiting Pretrained Model Reuse and Progressive Learning for Efficient Vision Transformer Scaling and Searching
Cheng Fu, Hanxian Huang, Zixuan Jiang, Yun Ni, Lifeng Nai, Gang Wu, Liqun Cheng, Yanqi Zhou, Sheng Li, Andrew Li, Jishen Zhao; Proceedings of the IEEE/CVF Inter

national Conference on Computer Vision (ICCV), 2023, pp. 17153-17163
One promising way to accelerate transformer training is to reuse small pretraine
d models to initialize the transformer, as their existing representation power f
acilitates faster model convergence. Previous works designed expansion operators
 to scale up pretrained models to the target model before training. Yet, model f
unctionality is difficult to preserve when scaling a transformer in all dimensio
ns at once. Moreover, maintaining the pretrained optimizer states for weights is
 critical for model scaling, whereas the new weights added during expansion lack
 these states in pretrained models. To address these issues, we propose TripLe,
which partially scales a model before training, while growing the rest of the ne
w parameters during training by copying both the warmed-up weights with the opti
mizer states from existing weights. As such, the new parameters introduced durin
g training will obtain their training states. Furthermore, through serializing t
he scaling of model width and depth, the functionality of each expansion can be
preserved. We evaluate TripLe in both single-trial model scaling and multi-trial
 neural architecture search (NAS). Due to the fast training convergence of TripL
e, the proxy accuracy from TripLe better reveals the model quality compared to f
rom-scratch training in multi-trial NAS. Experiments show that TripLe outperform
s both from-scratch training and knowledge distillation (KD) in both training ti
me and task performance. TripLe can also be combined with KD to achieve an even
higher task accuracy. For NAS, the model obtained from TripLe outperforms DeiT-B
 in task accuracy with 69% reduction in parameter size and FLOPs.
****
LogicSeg: Parsing Visual Semantics with Neural Logic Learning and Reasoning
Liulei Li, Wenguan Wang, Yi Yang; Proceedings of the IEEE/CVF International Conf
erence on Computer Vision (ICCV), 2023, pp. 4122-4133
Current high-performance semantic segmentation models are purely data-driven sub
-symbolic approaches and blind to the structured nature of the visual world. Thi
s is in stark contrast to human cognition which abstracts visual perceptions at
multiple levels and conducts symbolic reasoning with such structured abstraction
. To fill these fundamental gaps, we devise LogicSeg, a holistic visual semantic
 parser that integrates neural inductive learning and logic reasoning with both
rich data and symbolic knowledge. In particular, the semantic concepts of intere
st are structured as a hierarchy, from which a comprehensive set of constraints
are derived for describing the symbolic relations and formalized in first-order
logic. After fuzzy logic-based continuous relaxation, logical formulae are groun
ded onto data and neural computational graphs, hence enabling logic-induced netw
ork training. During inference, logical constraints are packaged into an iterati
ve process and injected into the network in a form of several matrix multiplicat
ions, so as to achieve hierarchy-coherent prediction with logic reasoning. These
 designs together make LogicSeg a general and compact neural-logic machine that
is readily integrated into existing segmentation models. Extensive experiments o
ver four datasets with various segmentation models and backbones verify the effe
ctiveness and generality of LogicSeg. We believe this study opens a new avenue f
or visual semantic parsing. Our code will be released.
****
The Devil is in the Upsampling: Architectural Decisions Made Simpler for Denoisi
ng with Deep Image Prior
Yilin Liu, Jiang Li, Yunkui Pang, Dong Nie, Pew-Thian Yap; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12408-12417
Deep Image Prior (DIP) shows that some network architectures inherently tend tow
ards generating smooth images while resisting noise, a phenomenon known as spect
ral bias. Image denoising is a natural application of this property. Although de
noising with DIP mitigates the need for large training sets, two often intertwin
ed practical challenges need to be overcome: architectural design and noise fitt
ing. Existing methods either handcraft or search for suitable architectures from
 a vast design space, due to the limited understanding of how architectural choi
ces affect the denoising outcome. In this study, we demonstrate from a frequency
 perspective that unlearnt upsampling is the main driving force behind the denoi
sing phenomenon with DIP. This finding leads to straightforward strategies for i

dentifying a suitable architecture for every image without laborious search. Extensive experiments show that the estimated architectures achieve superior denoising results than existing methods with up to 95% fewer parameters. Thanks to this under-parameterization, the resulting architectures are less prone to noise-fitting.
****
Video Object Segmentation-aware Video Frame Interpolation
Jun-Sang Yoo, Hongjae Lee, Seung-Won Jung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12322-12333
Video frame interpolation (VFI) is a very active research topic due to its broad applicability to many applications, including video enhancement, video encoding, and slow-motion effects. VFI methods have been advanced by improving the overall image quality for challenging sequences containing occlusions, large motion, and dynamic texture. This mainstream research direction neglects that foreground and background regions have different importance in perceptual image quality. Moreover, accurate synthesis of moving objects can be of utmost importance in computer vision applications. In this paper, we propose a video object segmentation (VOS)-aware training framework called VOS-VFI that allows VFI models to interpolate frames with more precise object boundaries. Specifically, we exploit VOS as an auxiliary task to help train VFI models by providing additional loss functions, including segmentation loss and bi-directional consistency loss. From extensive experiments, we demonstrate that VOS-VFI can boost the performance of existing VFI models by rendering clear object boundaries. Moreover, VOS-VFI displays its effectiveness on multiple benchmarks for different applications, including video object segmentation, object pose estimation, and visual tracking.
****
Coherent Event Guided Low-Light Video Enhancement
Jinxiu Liang, Yixin Yang, Boyu Li, Peiqi Duan, Yong Xu, Boxin Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10615-10625
With frame-based cameras, capturing fast-moving scenes without suffering from blur often comes at the cost of low SNR and low contrast. Worse still, the photometric constancy that enhancement techniques heavily relied on is fragile for frames with short exposure. Event cameras can record brightness changes at an extremely high temporal resolution. For low-light videos, event data are not only suitable to help capture temporal correspondences but also provide alternative observations in the form of intensity ratios between consecutive frames and exposure-invariant information. Motivated by this, we propose a low-light video enhancement method with hybrid inputs of events and frames. Specifically, a neural network is trained to establish spatiotemporal coherence between visual signals with different modalities and resolutions by constructing correlation volume across space and time. Experimental results on synthetic and real data demonstrate the superiority of the proposed method compared to the state-of-the-art methods.
****
Texture Learning Domain Randomization for Domain Generalized Segmentation
Sunghwan Kim, Dae-hwan Kim, Hoseong Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 677-687
Deep Neural Networks (DNNs)-based semantic segmentation models trained on a source domain often struggle to generalize to unseen target domains, i.e., a domain gap problem. Texture often contributes to the domain gap, making DNNs vulnerable to domain shift because they are prone to be texture-biased. Existing Domain Generalized Semantic Segmentation (DGSS) methods have alleviated the domain gap problem by guiding models to prioritize shape over texture. On the other hand, shape and texture are two prominent and complementary cues in semantic segmentation. This paper argues that leveraging texture is crucial for improving performance in DGSS. Specifically, we propose a novel framework, coined Texture Learning Domain Randomization (TLDR). TLDR includes two novel losses to effectively enhance texture learning in DGSS: (1) a texture regularization loss to prevent overfitting to source domain textures by using texture features from an ImageNet pre-trained model and (2) a texture generalization loss that utilizes random style imag

es to learn diverse texture representations in a self-supervised manner. Extensive experimental results demonstrate the superiority of the proposed TLDR; e.g., TLDR achieves 46.5 mIoU on GTA-to-Cityscapes using ResNet-50, which improves the prior state-of-the-art method by 1.9 mIoU. The source code is available at https://github.com/sssshwan/TLDR.
****
FCCNs: Fully Complex-valued Convolutional Networks using Complex-valued Color Model and Loss Function
Saurabh Yadav, Koteswar Rao Jerripothula; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10689-10698
Although complex-valued convolutional neural networks (iCNNs) have existed for a while, they lack proper complex-valued image inputs and loss functions. In addition, all their operations are not complex-valued as they have both complex-valued convolutional layers and real-valued fully-connected layers. As a result, they lack an end-to-end flow of complex-valued information, making them inconsistent w.r.t. the claimed operating domain, i.e., complex numbers. Considering these inconsistencies, we propose a complex-valued color model and loss function and turn fully-connected layers into convolutional layers. All these contributions culminate in what we call FCCNs (Fully Complex-valued Convolutional Networks), which take complex-valued images as inputs, perform only complex-valued operations, and have a complex-valued loss function. Thus, our proposed FCCNs have an end-to-end flow of complex-valued information, which lacks in existing iCNNs. Our extensive experiments on five image classification benchmark datasets show that FCCNs consistently perform better than existing iCNNs. Code is available at https://github.com/saurabhya/FCCNs .
****
Learning Concise and Descriptive Attributes for Visual Recognition
An Yan, Yu Wang, Yiwu Zhong, Chengyu Dong, Zexue He, Yujie Lu, William Yang Wang, Jingbo Shang, Julian McAuley; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3090-3100
Recent advances in foundation models present new opportunities for interpretable visual recognition -- one can first query Large Language Models (LLMs) to obtain a set of attributes that describe each class, then apply vision-language models to classify images via these attributes. Pioneering work shows that querying thousands of attributes can achieve performance competitive with image features. However, our further investigation on 8 datasets reveals that LLM-generated attributes in a large quantity perform almost the same as random words. This surprising finding suggests that significant noise may be present in these attributes. We hypothesize that there exist subsets of attributes that can maintain the classification performance with much smaller sizes, and propose a novel learning-to-search method to discover those concise sets of attributes. As a result, on the CUB dataset, our method achieves performance close to that of massive LLM-generated attributes (e.g., 10k attributes for CUB), yet using only 32 attributes in total to distinguish 200 bird species. Furthermore, our new paradigm demonstrates several additional benefits: higher interpretability and interactivity for humans, and the ability to summarize knowledge for a recognition task.
****
Learning Unified Decompositional and Compositional NeRF for Editable Novel View Synthesis
Yuxin Wang, Wayne Wu, Dan Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18247-18256
Implicit neural representations have shown powerful capacity in modeling real-world 3D scenes, offering superior performance in novel view synthesis. In this paper, we target a more challenging scenario, i.e., joint scene novel view synthesis and editing based on implicit neural scene representations. State-of-the-art methods in this direction typically consider building separate networks for these two tasks (i.e., view synthesis and editing). Thus, the modeling of interactions and correlations between these two tasks is very limited, which, however, is critical for learning high-quality scene representations.To tackle this problem, in this paper, we propose a unified Neural Radiance Field (NeRF) framework to e

ffectively perform joint scene decomposition and composition for modeling real-world scenes. The decomposition aims at learning disentangled 3D representations of different objects and the background, allowing for scene editing, while scene composition models an entire scene representation for novel view synthesis. Specifically, with a two-stage NeRF framework, we learn a coarse stage for predicting a global radiance field as guidance for point sampling, and in the second fine-grained stage, we perform scene decomposition by a novel one-hot object radiance field regularization module and a pseudo supervision via inpainting to handle ambiguous background regions occluded by objects. The decomposed object-level radiance fields are further composed by using activations from the decomposition module. Extensive quantitative and qualitative results show the effectiveness of our method for scene decomposition and composition, outperforming state-of-the-art methods for both novel-view synthesis and editing tasks.
****

Label-Noise Learning with Intrinsically Long-Tailed Data

Yang Lu, Yiliang Zhang, Bo Han, Yiu-ming Cheung, Hanzi Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1369-1378

Label noise is one of the key factors that lead to the poor generalization of deep learning models. Existing label-noise learning methods usually assume that the ground-truth classes of the training data are balanced. However, the real-world data is often imbalanced, leading to the inconsistency between observed and intrinsic class distribution with label noises. In this case, it is hard to distinguish clean samples from noisy samples on the intrinsic tail classes with the unknown intrinsic class distribution. In this paper, we propose a learning framework for label-noise learning with intrinsically long-tailed data. Specifically, we propose two-stage bi-dimensional sample selection (TABASCO) to better separate clean samples from noisy samples, especially for the tail classes. TABASCO consists of two new separation metrics that complement each other to compensate for the limitation of using a single metric in sample separation. Extensive experiments on benchmarks demonstrate the effectiveness of our method. Our code is available at https://github.com/Wakings/TABASCO.
****

SeeABLE: Soft Discrepancies and Bounded Contrastive Learning for Exposing Deepfakes

Nicolas Larue, Ngoc-Son Vu, Vitomir Struc, Peter Peer, Vassilis Christophides; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21011-21021

Modern deepfake detectors have achieved encouraging results, when training and test images are drawn from the same data collection. However, when these detectors are applied to images produced with unknown deepfake-generation techniques, considerable performance degradations are commonly observed. In this paper, we propose a novel deepfake detector, called SeeABLE, that formalizes the detection problem as a (one-class) out-of-distribution detection task and generalizes better to unseen deepfakes. Specifically, SeeABLE first generates local image perturbations (referred to as soft-discrepancies) and then pushes the perturbed faces towards predefined prototypes using a novel regression-based bounded contrastive loss. To strengthen the generalization performance of SeeABLE to unknown deepfake types, we generate a rich set of soft discrepancies and train the detector: (i) to localize, which part of the face was modified, and (ii) to identify the alteration type. To demonstrate the capabilities of SeeABLE, we perform rigorous experiments on several widely-used deepfake datasets and show that our model convincingly outperforms competing state-of-the-art detectors, while exhibiting highly encouraging generalization capabilities. The source code for SeeABLE is available from: https://github.com/anonymous-author-sub/seeable.
****

Semi-Supervised Learning via Weight-Aware Distillation under Class Distribution Mismatch

Pan Du, Suyun Zhao, Zisen Sheng, Cuiping Li, Hong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16410-16420

Semi-Supervised Learning (SSL) under class distribution mismatch aims to tackle a challenging problem wherein unlabeled data contain lots of unknown categories unseen in the labeled ones. In such mismatch scenarios, traditional SSL suffers severe performance damage due to the harmful invasion of the instances with unknown categories into the target classifier. In this study, by strict mathematical reasoning, we reveal that the SSL error under class distribution mismatch is composed of pseudo-labeling error and invasion error, both of which jointly bound the SSL population risk. To alleviate the SSL error, we propose a robust SSL framework called Weight-Aware Distillation (WAD) that, by weights, selectively transfers knowledge beneficial to the target task from unsupervised contrastive representation to the target classifier. Specifically, WAD captures adaptive weights and high-quality pseudo-labels to target instances by exploring point mutual information (PMI) in representation space to maximize the role of unlabeled data and filter unknown categories. Theoretically, we prove that WAD has a tight upper bound of population risk under class distribution mismatch. Experimentally, extensive results demonstrate that WAD outperforms five state-of-the-art SSL approaches

and one standard baseline on two benchmark datasets, CIFAR10 and CIFAR100, and an artificial cross-dataset. The code is available at https://github.com/RUC-DWBI-ML/research/tree/main/WAD-master.
****

ELFNet: Evidential Local-global Fusion for Stereo Matching
Jieming Lou, Weide Liu, Zhuo Chen, Fayao Liu, Jun Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17784-17793
Although existing stereo matching models have achieved continuous improvement, they often face issues related to trustworthiness due to the absence of uncertainty estimation. Additionally, effectively leveraging multi-scale and multi-view knowledge of stereo pairs remains unexplored. In this paper, we introduce the Evidential Local-global Fusion (ELF) framework for stereo matching, which endows both uncertainty estimation and confidence-aware fusion with trustworthy heads. Instead of predicting the disparity map alone, our model estimates an evidential-based disparity considering both aleatoric and epistemic uncertainties. With the normal inverse-Gamma distribution as a bridge, the proposed framework realizes intra evidential fusion of multi-level predictions and inter evidential fusion between cost-volume-based and transformer-based stereo matching. Extensive experimental results show that the proposed framework exploits multi-view information effectively and achieves state-of-the-art overall performance both on accuracy and cross-domain generalization. The codes are available at https://github.com/jimmy19991222/ELFNet.
****

SimpleClick: Interactive Image Segmentation with Simple Vision Transformers
Qin Liu, Zhenlin Xu, Gedas Bertasius, Marc Niethammer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22290-22300
Click-based interactive image segmentation aims at extracting objects with a limited user clicking. A hierarchical backbone is the de-facto architecture for current methods. Recently, the plain, non-hierarchical Vision Transformer (ViT) has emerged as a competitive backbone for dense prediction tasks. This design allows the original ViT to be a foundation model that can be finetuned for downstream tasks without redesigning a hierarchical backbone for pretraining. Although this design is simple and has been proven effective, it has not yet been explored for interactive segmentation. To fill this gap, we propose SimpleClick, the first plain-backbone method for interactive segmentation. Other than the plain backbone, we also explore several variants of simple feature pyramid networks that only take as input the last feature representation of the backbone. With the plain backbone pretrained as a masked autoencoder (MAE), SimpleClick achieves state-of-the-art performance. Remarkably, our method achieves 4.15 NoC@90 on SBD, improving 21.8% over the previous best result. Extensive evaluation on medical images demonstrates the generalizability of our method. We further develop an extremely tiny ViT backbone for SimpleClick and provide a detailed computational analysis, highlighting its suitability as a practical annotation tool.
****

Towards Content-based Pixel Retrieval in Revisited Oxford and Paris
Guoyuan An, Woo Jae Kim, Saelyne Yang, Rong Li, Yuchi Huo, Sun-Eui Yoon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20507-20518
This paper introduces the first two landmark pixel retrieval benchmarks. Like semantic segmentation extends classification to the pixel level, pixel retrieval is an extension of image retrieval and offers information about which pixels are related to the query object. In addition to retrieving images for the given query, it helps users quickly identify the query object in true positive images and exclude false positive images by denoting the correlated pixels. Our user study results show pixel-level annotation can significantly improve the user experience. Compared with semantic and instance segmentation, pixel retrieval requires a fine-grained recognition capability for variable-granularity targets. To this end, we propose pixel retrieval benchmarks named PROxford and PRParis, which are based on the widely used image retrieval datasets, ROxford and RParis. Three professional annotators label 5,942 images with two rounds of double-checking and refinement. Furthermore, we conduct extensive experiments and analysis on the SOTA methods in image search, image matching, detection, segmentation, and dense matching using our pixel retrieval benchmarks. Results show that the pixel retrieval task is challenging to these approaches and distinctive from existing problems, suggesting that further research can advance the content-based pixel-retrieval and, thus, user search experience.
****
S-TREK: Sequential Translation and Rotation Equivariant Keypoints for Local Feature Extraction
Emanuele Santellani, Christian Sormann, Mattia Rossi, Andreas Kuhn, Friedrich Fraundorfer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9728-9737
In this work we introduce S-TREK, a novel local feature extractor that combines a deep keypoint detector, which is both translation and rotation equivariant by design, with a lightweight deep descriptor extractor. We train the S-TREK keypoint detector within a framework inspired by reinforcement learning, where we leverage a sequential procedure to maximize a reward directly related to keypoint repeatability. Our descriptor network is trained following a "detect, then describe" approach, where the descriptor loss is evaluated only at those locations where keypoints have been selected by the already trained detector. Extensive experiments on multiple benchmarks confirm the effectiveness of our proposed method, with S-TREK often outperforming other state-of-the-art methods in terms of repeatability and quality of the recovered poses, especially when dealing with in-plane rotations.
****
Retro-FPN: Retrospective Feature Pyramid Network for Point Cloud Semantic Segmentation
Peng Xiang, Xin Wen, Yu-Shen Liu, Hui Zhang, Yi Fang, Zhizhong Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17826-17838
Learning per-point semantic features from the hierarchical feature pyramid is essential for point cloud semantic segmentation. However, most previous methods suffered from ambiguous region features or failed to refine per-point features effectively, which leads to information loss and ambiguous semantic identification. To resolve this, we propose Retro-FPN to model the per-point feature prediction as an explicit and retrospective refining process, which goes through all the pyramid layers to extract semantic features explicitly for each point. Its key novelty is a retro-transformer for summarizing semantic contexts from the previous layer and accordingly refining the features in the current stage. In this way, the categorization of each point is conditioned on its local semantic pattern. Specifically, the retro-transformer consists of a local cross-attention block and a semantic gate unit. The cross-attention serves to summarize the semantic pattern retrospectively from the previous layer. And the gate unit carefully incorporates the summarized contexts and refines the current semantic features. Retro-F

PN is a pluggable neural network that applies to hierarchical decoders. By integ rating Retro-FPN with three representative backbones, including both point-based and voxel-based methods, we show that Retro-FPN can significantly improve perfo rmance over state-of-the-art backbones. Comprehensive experiments on widely used benchmarks can justify the effectiveness of our design. The source is available at https://github.com/AllenXiangX/Retro-FPN.
****
Rethinking Range View Representation for LiDAR Segmentation
Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 228-240
LiDAR segmentation is crucial for autonomous driving perception. Recent trends favor point- or voxel-based methods as they often yield better performance than the traditional range view representation. In this work, we unveil several key factors in building powerful range view models. We observe that the "many-to-one" mapping, semantic incoherence, and shape deformation are possible impediments against effective learning from range view projections. We present RangeFormer -- a full-cycle framework comprising novel designs across network architecture, data augmentation, and post-processing -- that better handles the learning and processing of LiDAR point clouds from the range view. We further introduce a Scalable Training from Range view (STR) strategy that trains on arbitrary low-resolution 2D range images, while still maintaining satisfactory 3D segmentation accuracy. We show that, for the first time, a range view method is able to surpass the point, voxel, and multi-view fusion counterparts in the competing LiDAR semantic and panoptic segmentation benchmarks, i.e., SemanticKITTI, nuScenes, and ScribbleKITTI.
****
Divide and Conquer: 3D Point Cloud Instance Segmentation With Point-Wise Binarization
Weiguang Zhao, Yuyao Yan, Chaolong Yang, Jianan Ye, Xi Yang, Kaizhu Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 562-571
Instance segmentation on point clouds is crucially important for 3D scene understanding. Most SOTAs adopt distance clustering, which is typically effective but does not perform well in segmenting adjacent objects with the same semantic label (especially when they share neighboring points). Due to the uneven distribution of offset points, these existing methods can hardly cluster all instance points. To this end, we design a novel divide-and-conquer strategy named PBNet that binarizes each point and clusters them separately to segment instances. Our binary clustering divides offset instance points into two categories: high and low density points (HPs vs. LPs). Adjacent objects can be clearly separated by removing LPs, and then be completed and refined by assigning LPs via a neighbor voting method. To suppress potential over-segmentation, we propose to construct local scenes with the weight mask for each instance. As a plug-in, the proposed binary clustering can replace the traditional distance clustering and lead to consistent performance gains on many mainstream baselines. A series of experiments on ScanNetV2 and S3DIS datasets indicate the superiority of our model. In particular, PBNet ranks first on the ScanNetV2 official benchmark challenge, achieving the highest mAP. Code will be available publicly at https://github.com/weiguangzhao/PBNet.
****
BANSAC: A Dynamic BAyesian Network for Adaptive SAmple Consensus
Valter Piedade, Pedro Miraldo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3738-3747
RANSAC-based algorithms are the standard techniques for robust estimation in computer vision. These algorithms are iterative and computationally expensive; they alternate between random sampling of data, computing hypotheses, and running inlier counting. Many authors tried different approaches to improve efficiency. One of the major improvements is having a guided sampling, letting the RANSAC cycle stop sooner. This paper presents a new adaptive sampling process for RANSAC. P

revious methods either assume no prior information about the inlier/outlier classification of data points or use some previously computed scores in the sampling. In this paper, we derive a dynamic Bayesian network that updates individual data points' inlier scores while iterating RANSAC. At each iteration, we apply weighted sampling using the updated scores. Our method works with or without prior data point scorings. In addition, we use the updated inlier/outlier scoring for deriving a new stopping criterion for the RANSAC loop. We test our method in multiple real-world datasets for several applications and obtain state-of-the-art results. Our method outperforms the baselines in accuracy while needing less computational time.
****
ShapeScaffolder: Structure-Aware 3D Shape Generation from Text
Xi Tian, Yong-Liang Yang, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2715-2724
We present ShapeScaffolder, a structure-based neural network for generating colored 3D shapes based on text input. The approach, similar to providing scaffolds as internal structural supports and adding more details to them, aims to capture finer text-shape connections and improve the quality of generated shapes. Traditional text-to-shape methods often generate 3D shapes as a whole. However, humans tend to understand both shape and text as being structure-based. For example, a table is interpreted as being composed of legs, a seat, and a back; similarly, texts possess inherent linguistic structures that can be analyzed as dependency graphs, depicting the relationships between entities within the text. We believe structure-aware shape generation can bring finer text-shape connections and improve shape generation quality. However, the lack of explicit shape structure and the high freedom of text structure make cross-modality learning challenging. To address these challenges, we first build the structured shape implicit fields in an unsupervised manner. We then propose the part-level attention mechanism between shape parts and textual graph nodes to align the two modalities at the structural level. Finally, we employ a shape refiner to add further detail to the predicted structure, yielding the final results. Extensive experimentation demonstrates that our approaches outperform state-of-the-art methods in terms of both shape fidelity and shape-text matching. Our methods also allow for part-level manipulation and improved part-level completeness.
****
Read-only Prompt Optimization for Vision-Language Few-shot Learning
Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, Hyunwoo J. Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1401-1411
In recent years, prompt tuning has proven effective in adapting pre-trained vision-language models to down- stream tasks. These methods aim to adapt the pre-trained models by introducing learnable prompts while keeping pre- trained weights frozen. However, learnable prompts can affect the internal representation within the self-attention module, which may negatively impact performance vari- ance and generalization, especially in data-deficient set- tings. To address these issues, we propose a novel ap- proach, Read-only Prompt Optimization (RPO). RPO lever- ages masked attention to prevent the internal representa- tion shift in the pre-trained model. Further, to facilitate the optimization of RPO, the read-only prompts are ini- tialized based on special tokens of the pre-trained model. Our extensive experiments demonstrate that RPO outper- forms CLIP and CoCoOp in base-to-new generalization and domain generalization while displaying better robust- ness. Also, the proposed method achieves better generaliza- tion on extremely data-deficient settings, while improving parameter efficiency and computational overhead. Code is
available at https://github.com/mlvlab/RPO.
****
COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts
Xiaofeng Mao, Yuefeng Chen, Yao Zhu, Da Chen, Hang Su, Rong Zhang, Hui Xue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6339-6350

Practical object detection application can lose its effectiveness on image input s with natural distribution shifts. This problem leads the research community to pay more attention on the robustness of detectors under Out-Of-Distribution (OO D) inputs. Existing works construct datasets to benchmark the detector's OOD rob ustness for a specific application scenario, e.g., Autonomous Driving. However, these datasets lack universality and are hard to benchmark general detectors bui lt on common tasks such as COCO. To give a more comprehensive robustness assessm ent, we introduce COCO-O(ut-of-distribution), a test dataset based on COCO with 6 types of natural distribution shifts. COCO-O has a large distribution gap with training data and results in a significant 55.7% relative performance drop on a Faster R-CNN detector. We leverage COCO-O to conduct experiments on more than 1 00 modern object detectors to investigate if their improvements are credible or just over-fitting to the COCO test set. Unfortunately, most classic detectors in early years do not exhibit strong OOD generalization. We further study the robu stness effect on recent breakthroughs of detector's architecture design, augment ation and pre-training techniques. Some empirical findings are revealed: 1) Comp ared with detection head or neck, backbone is the most important part for robust ness; 2) An end-to-end detection transformer design brings no enhancement, and m ay even reduce robustness; 3) Large-scale foundation models have made a great le ap on robust object detection. We hope our COCO-O could provide a rich testbed f or robustness study of object detection. The dataset will be available at https: //github.com/alibaba/easyrobust/tree/main/benchmarks/coco_o.
****
E2NeRF: Event Enhanced Neural Radiance Fields from Blurry Images
Yunshan Qi, Lin Zhu, Yu Zhang, Jia Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13254-13264
Neural Radiance Fields (NeRF) achieves impressive ren-dering performance by lear ning volumetric 3D representation from several images of different views. Howeve r, it is difficult to reconstruct a sharp NeRF from blurry input as often occurr ed in the wild. To solve this problem, we propose a novel Event-Enhanced NeRF (E 2NeRF) by utilizing the combination data of a bio-inspired event camera and a st andard RGB camera. To effectively introduce event stream into the learning proce ss of neural volumetric representation, we propose a blur rendering loss and an event rendering loss, which guide the network via modelling real blur process an d event generation process, respectively. Moreover, a camera pose estimation fra mework for real-world data is built with the guidance of event stream to general ize the method to practical applications. In contrast to previous image-based or event-based NeRF, our framework effectively utilizes the internal relationship between events and images. As a result, E2NeRF not only achieves image deblurrin g but also achieves high-quality novel view image generation. Extensive experime nts on both synthetic data and real-world data demonstrate that E2NeRF can effec tively learn a sharp NeRF from blurry images, especially in complex and low-ligh t scenes. Our code and datasets are publicly available at https://github.com/iCV TEAM/E2NeRF.
****
EgoTV: Egocentric Task Verification from Natural Language Task Descriptions
Rishi Hazra, Brian Chen, Akshara Rai, Nitin Kamra, Ruta Desai; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15417-1 5429
To enable progress towards egocentric agents capable of understanding everyday t asks specified in natural language, we propose a benchmark and a synthetic datas et called Egocentric Task Verification (EgoTV). The goal in EgoTV is to verify t he execution of tasks from egocentric videos based on the natural language descr iption of these tasks. EgoTV contains pairs of videos and their task description s for multi-step tasks -- these tasks contain multiple sub-task decompositions, state changes, object interactions, and sub-task ordering constraints. In additi on, EgoTV also provides abstracted task descriptions that contain only partial d etails about ways to accomplish a task. Consequently, EgoTV requires causal, tem poral, and compositional reasoning of video and language modalities, which is mi ssing in existing datasets. We also find that existing vision-language models st

ruggle at such all round reasoning needed for task verification in EgoTV. Inspired by the needs of EgoTV, we propose a novel Neuro-Symbolic Grounding (NSG) approach that leverages symbolic representations to capture the compositional and temporal structure of tasks. We demonstrate NSG's capability towards task tracking and verification on our EgoTV dataset and a real-world dataset derived from CrossTask (CTV). We open-source the EgoTV and CTV datasets and the NSG model for future research on egocentric assistive agents.
****
Benchmarking Low-Shot Robustness to Natural Distribution Shifts
Aaditya Singh, Kartik Sarangmath, Prithvijit Chattopadhyay, Judy Hoffman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16232-16242
Robustness to natural distribution shifts has seen remarkable progress thanks to recent pre-training strategies combined with better fine-tuning methods. However, such fine-tuning assumes access to large amounts of labelled data, and the extent to which the observations hold when the amount of training data is not as high remains unknown. We address this gap by performing the first in-depth study of robustness to various natural distribution shifts in different low-shot regimes: spanning datasets, architectures, pre-trained initializations, and state-of-the-art robustness interventions. Most importantly, we find that there is no single model of choice that is often more robust than others, and existing interventions can fail to improve robustness on some datasets even if they do so in the full-shot regime. We hope that our work will motivate the community to focus on this problem of practical importance.
****
AdaptGuard: Defending Against Universal Attacks for Model Adaptation
Lijun Sheng, Jian Liang, Ran He, Zilei Wang, Tieniu Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19093-19103
Model adaptation aims at solving the domain transfer problem under the constraint of only accessing the pretrained source models. With the increasing considerations of data privacy and transmission efficiency, this paradigm has been gaining recent popularity. This paper studies the vulnerability to universal attacks transferred from the source domain during model adaptation algorithms due to the existence of malicious providers. We explore both universal adversarial perturbations and backdoor attacks as loopholes on the source side and discover that they still survive in the target models after adaptation. To address this issue, we propose a model preprocessing framework, named AdaptGuard, to improve the security of model adaptation algorithms. AdaptGuard avoids direct use of the risky source parameters through knowledge distillation and utilizes the pseudo adversarial samples under adjusted radius to enhance the robustness. AdaptGuard is a plug-and-play module that requires neither robust pretrained models nor any changes for the following model adaptation algorithms. Extensive results on three commonly used datasets and two popular adaptation methods validate that AdaptGuard can effectively defend against universal attacks and maintain clean accuracy in the target domain simultaneously. We hope this research will shed light on the safety and robustness of transfer learning. Code is available at https://github.com/TomSheng21/AdaptGuard.
****
StageInteractor: Query-based Object Detector with Cross-stage Interaction
Yao Teng, Haisong Liu, Sheng Guo, Limin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6577-6588
Previous object detectors make predictions based on dense grid points or numerous preset anchors. Most of these detectors are trained with one-to-many label assignment strategies. On the contrary, recent query-based object detectors are based a sparse set of learnable queries refined by a series of decoder layers. The one-to-one label assignment is independently applied on each layer for deep supervision during training. Despite the great success of query-based object detection, however, this vanilla one-to-one label assignment strategy requires the detectors to have strong fine-grained discrimination and modeling capacity. In this paper, we propose a new query-based object detector with cross-stage interaction

, coined as StageInteractor. During the forward pass, we come up with an efficient way to improve this modeling ability by reusing dynamic operators with lightweight adapters. As for the label assignment, a cross-stage label assigner is designed to improve the one-to-one label assignment. With this assigner, the training target class labels are gathered across stages and then reallocated to proper predictions at each decoder layer. On MS COCO benchmark, our model improves the baseline counterpart by 2.2 AP, and achieves a 44.8 AP with ResNet-50 as backbone, 100 queries and 12 training epochs. With longer training time and 300 queries, StageInteractor achieves 51.3 AP and 52.7 AP with ResNeXt-101-DCN and Swin-S, respectively. The code and models are made available at https://github.com/MCG-NJU/StageInteractor.
****
DeLiRa: Self-Supervised Depth, Light, and Radiance Fields
Vitor Guizilini, Igor Vasiljevic, Jiading Fang, Rares Ambrus, Sergey Zakharov, Vincent Sitzmann, Adrien Gaidon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17935-17945
Differentiable volumetric rendering is a powerful paradigm for 3D reconstruction and novel view synthesis. However, standard volume rendering approaches struggle with degenerate geometries in the case of limited viewpoint diversity, a common scenario in robotics applications. In this work, we propose to use the multi-view photometric objective from the self-supervised depth estimation literature as a geometric regularizer for volumetric rendering, significantly improving novel view synthesis without requiring additional information. Building upon this insight, we explore the explicit modeling of scene geometry using a generalist Transformer, jointly learning a radiance field as well as depth and light fields with a set of shared latent codes. We demonstrate that sharing geometric information across tasks is mutually beneficial, leading to improvements over single-task learning without an increase in network complexity. Our DeLiRa architecture achieves state-of-the-art results on the ScanNet benchmark, enabling high quality volumetric rendering as well as real-time novel view and depth synthesis in the limited viewpoint diversity setting.
****
Moment Detection in Long Tutorial Videos
Ioana Croitoru, Simion-Vlad Bogolin, Samuel Albanie, Yang Liu, Zhaowen Wang, Seunghyun Yoon, Franck Dernoncourt, Hailin Jin, Trung Bui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2594-2604
Tutorial videos play an increasingly important role in professional development and self-directed education. For users to realise the full benefits of this medium, tutorial videos must be efficiently searchable. In this work, we focus on the task of moment detection, in which the goal is to localise the temporal window where a given event occurs within a given tutorial video. Prior work on moment detection has focused primarily on short videos (typically on videos shorter than three minutes). However, many tutorial videos are substantially longer (stretching to hours in duration), presenting significant challenges for existing moment detection approaches. To study this problem, we propose the first dataset of untrimmed, long-form tutorial videos for the task of Moment Detection called the Behance Moment Detection (BMD) dataset. BMD videos have an average duration of over one hour and are characterised by slowly evolving visual content and wide-ranging dialogue. To meet the unique challenges of this dataset, we propose a new framework, LongMoment-DETR, and demonstrate that it outperforms strong baselines. Additionally, we introduce a variation of the dataset that contains YouTube Chapter annotations and show that the features obtained by our framework can be successfully used to boost the performance on the task of chapter detection. Code and data can be found at https://github.com/ioanacroi/longmoment-detr.
****
Stable Cluster Discrimination for Deep Clustering
Qi Qian; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16645-16654
Deep clustering can optimize representations of instances (i.e., representation learning) and explore the inherent data distribution (i.e., clustering) simultan

eously, which demonstrates a superior performance over conventional clustering methods with given features. However, the coupled objective implies a trivial solution that all instances collapse to the uniform features. To tackle the challenge, a two-stage training strategy is developed for decoupling, where it introduces an additional pre-training stage for representation learning and then fine-tunes the obtained model for clustering. Meanwhile, one-stage methods are developed mainly for representation learning rather than clustering, where various constraints for cluster assignments are designed to avoid collapsing explicitly. Despite the success of these methods, an appropriate learning objective tailored for deep clustering has not been investigated sufficiently. In this work, we first show that the prevalent discrimination task in supervised learning is unstable for one-stage clustering due to the lack of ground-truth labels and positive instances for certain clusters in each mini-batch. To mitigate the issue, a novel stable cluster discrimination (SeCu) task is proposed and a new hardness-aware clustering criterion can be obtained accordingly. Moreover, a global entropy constraint for cluster assignments is studied with efficient optimization. Extensive experiments are conducted on benchmark data sets and ImageNet. SeCu achieves state-of-the-art performance on all of them, which demonstrates the effectiveness of one-stage deep clustering.
****
Pix2Video: Video Editing using Image Diffusion
Duygu Ceylan, Chun-Hao P. Huang, Niloy J. Mitra; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23206-23217
Image diffusion models, trained on massive image collections, have emerged as the most versatile image generator model in terms of quality and diversity. They support inverting real images and conditional (e.g., text) generation, making them attractive for high-quality image editing applications. We investigate how to use such pre-trained image models for text-guided video editing. The critical challenge is to achieve the target edits while still preserving the content of the source video. Our method works in two simple steps: first, we use a pre-trained structure-guided (e.g., depth) image diffusion model to perform text-guided edits on an anchor frame; then, in the key step, we progressively propagate the changes to the future frames via self-attention feature injection to adapt the core denoising step of the diffusion model. We then consolidate the changes by adjusting the latent code for the frame before continuing the process. Our approach is training-free and generalizes to a wide range of edits. We demonstrate the effectiveness of the approach by extensive experimentation and compare it against four different prior and parallel efforts (on ArXiv).
We demonstrate that realistic text-guided video edits are possible, without any compute-intensive preprocessing or video-specific finetuning.
****
DFA3D: 3D Deformable Attention For 2D-to-3D Feature Lifting
Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6684-6693
In this paper, we propose a new operator, called 3D DeFormable Attention (DFA3D), for 2D-to-3D feature lifting, which transforms multi-view 2D image features into a unified 3D space for 3D object detection. Existing feature lifting approaches, such as Lift-Splat-based and 2D attention-based, either use estimated depth to get pseudo LiDAR features and then splat them to a 3D space, which is a one-pass operation without feature refinement, or ignore depth and lift features by 2D attention mechanisms, which achieve finer semantics while suffering from a depth ambiguity problem. In contrast, our DFA3D-based method first leverages the estimated depth to expand each view's 2D feature map to 3D and then utilizes DFA3D to aggregate features from the expanded 3D feature maps. With the help of DFA3D, the depth ambiguity problem can be effectively alleviated from the root, and the lifted features can be progressively refined layer by layer, thanks to the Transformer-like architecture. In addition, we propose a mathematically equivalent implementation of DFA3D which can significantly improve its memory efficiency and computational speed. We integrate DFA3D into several methods that use 2D atte

ntion-based feature lifting with only a few modifications in code and evaluate on the nuScenes dataset. The experiment results show a consistent improvement of +1.41% mAP on average, and up to +15.1% mAP improvement when high-quality depth information is available, demonstrating the superiority, applicability, and huge potential of DFA3D. The code is available at https://github.com/IDEA-Research/3D-deformable-attention.git.
****
Holistic Geometric Feature Learning for Structured Reconstruction
Ziqiong Lu, Linxi Huan, Qiyuan Ma, Xianwei Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21807-21817
The inference of topological principles is a key problem in structured reconstruction. We observe that wrongly predicted topological relationships are often incurred by the lack of holistic geometry clues in low-level features. Inspired by the fact that massive signals can be compactly described with frequency analysis, we experimentally explore the efficiency and tendency of learning structure geometry in the frequency domain. Accordingly, we propose a frequency-domain feature learning strategy (F-Learn) to fuse scattered geometric fragments holistically for topology-intact structure reasoning. Benefiting from the parsimonious design, the F-Learn strategy can be easily deployed into a deep reconstructor with a lightweight model modification. Experiments demonstrate that the F-Learn strategy can effectively introduce structure awareness into geometric primitive detection and topology inference, bringing significant performance improvement to final structured reconstruction. Code and pre-trained models are available at https://github.com/Geo-Tell/F-Learn.
****
FateZero: Fusing Attentions for Zero-shot Text-based Video Editing
Chenyang QI, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, Qifeng Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15932-15942
The diffusion-based generative models have achieved remarkable success in text-based image generation. However, since it contains enormous randomness in generation progress, it is still challenging to apply such models for real-world visual content editing, especially in videos. In this paper, we propose FateZero, a zero-shot text-based editing method on real-world videos without per-prompt training or use-specific mask. To edit videos consistently, we propose several techniques based on the pre-trained models. Firstly, in contrast to the straightforward DDIM inversion technique, our approach captures intermediate attention maps during inversion, which effectively retain both structural and motion information. These maps are directly fused in the editing process rather than generated during denoising. To further minimize semantic leakage of the source video, we then fuse self-attentions with a blending mask obtained by cross-attention features from the source prompt. Furthermore, we have implemented a reform of the self-attention mechanism in denoising UNet by introducing spatial-temporal attention to ensure frame consistency. Yet succinct, our method is the first one to show the ability of zero-shot text-driven video style and local attribute editing from the trained text-to-image model. We also have a better zero-shot shape-aware editing ability in the text-to-video model.
****
Uncertainty-guided Learning for Improving Image Manipulation Detection
Kaixiang Ji, Feng Chen, Xin Guo, Yadong Xu, Jian Wang, Jingdong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22456-22465
Image manipulation detection (IMD) is of vital importance as faking images and spreading misinformation can be malicious and harm our daily life. IMD is the core technique to solve these issues and poses challenges in two main aspects: (1) Data Uncertainty, i.e., the manipulated artifacts are often hard for humans to discern and lead to noisy labels, which may disturb model training; (2) Model Uncertainty, i.e., the same object may hold different categories (tampered or not) due to manipulation operations, which could potentially confuse the model training and result in unreliable outcomes. Previous works mainly focus on solving the

 model uncertainty issue by designing meticulous features and networks, however, the data uncertainty problem is rarely considered. In this paper, we address both problems by introducing an uncertainty-guided learning framework, which measures data and model uncertainty by a novel Uncertainty Estimation Network (UEN). UEN is trained under dynamic supervision, and outputs estimated uncertainty maps to refine manipulation detection results, which significantly alleviates the learning difficulties. To our knowledge, this is the first work to embed uncertainty modeling into IMD. Extensive experiments on various datasets demonstrate state-of-the-art performance, validating the effectiveness and generalizability of our method.
****
LMR: A Large-Scale Multi-Reference Dataset for Reference-Based Super-Resolution
Lin Zhang, Xin Li, Dongliang He, Fu Li, Errui Ding, Zhaoxiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13118-13127
It is widely agreed that reference-based super-resolution (RefSR) achieves superior results by referring to similar high quality images, compared to single image super-resolution (SISR). Intuitively, the more references, the better performance. However, previous RefSR methods have all focused on single-reference image training, while multiple reference images are often available in testing or practical applications. The root cause of such training-testing mismatch is the absence of publicly available multi-reference SR training datasets, which greatly hinders research efforts on multi-reference super-resolution. To this end, we construct a large-scale, multi-reference super-resolution dataset, named LMR. It contains 112,142 groups of 300x300 training images, which is 10x of the existing largest RefSR dataset. The image size is also some times larger. More importantly, each group is equipped with 5 reference images with different similarity levels. Furthermore, we propose a new baseline method for multi-reference super-resolution: MRefSR, including a Multi-Reference Attention Module (MAM) for feature fusion of an arbitrary number of reference images, and a Spatial Aware Filtering Module (SAFM) for the fused feature selection. The proposed MRefSR achieves significant improvements over state-of-the-art approaches on both quantitative and qualitative evaluations. Our code and data are available at: https://github.com/wdmwhh/MRefSR.
****
Neural Implicit Surface Evolution
Tiago Novello, Vinicius da Silva, Guilherme Schardong, Luiz Schirmer, Helio Lopes, Luiz Velho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14279-14289
This work investigates the use of smooth neural networks for modeling dynamic variations of implicit surfaces under the level set equation (LSE). For this, it extends the representation of neural implicit surfaces to the space-time, which opens up mechanisms for continuous geometric transformations. Examples include evolving an initial surface towards general vector fields, smoothing and sharpening using the mean curvature equation, and interpolations of initial conditions. The network training considers two constraints. A data term is responsible for fitting the initial condition to the corresponding time instant. Then, a LSE term forces the network to approximate the underlying geometric evolution given by the LSE, without any supervision. The network can also be initialized based on previously trained initial conditions, resulting in faster convergence compared to the standard approach.
****
Distribution-Aligned Diffusion for Human Mesh Recovery
Lin Geng Foo, Jia Gong, Hossein Rahmani, Jun Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9221-9232
Recovering a 3D human mesh from a single RGB image is a challenging task due to depth ambiguity and self-occlusion, resulting in a high degree of uncertainty. Meanwhile, diffusion models have recently seen much success in generating high-quality outputs by progressively denoising noisy inputs. Inspired by their capability, we explore a diffusion-based approach for human mesh recovery, and propose

a Human Mesh Diffusion (HMDiff) framework which frames mesh recovery as a revers e diffusion process. We also propose a Distribution Alignment Technique (DAT) th at injects input-specific distribution information into the diffusion process, a nd provides useful prior knowledge to simplify the mesh recovery task. Our metho d achieves state-of-the-art performance on three widely used datasets. Project p age: https://gongjia0208.github.io/HMDiff/.
****
Rosetta Neurons: Mining the Common Units in a Model Zoo
Amil Dravid, Yossi Gandelsman, Alexei A. Efros, Assaf Shocher; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1934-19 43
Do different neural networks, trained for various vision tasks, share some commo n representations? In this paper, we demonstrate the existence of common feature s we call "Rosetta Neurons" across a range of models with different architecture s, different tasks (generative and discriminative), and different types of super vision (class-supervised, text-supervised, self-supervised). We present an algor ithm for mining a dictionary of Rosetta Neurons across several popular vision mo dels: Class Supervised-ResNet50, DINO-ResNet50, DINO-ViT, MAE, CLIP-ResNet50, Bi gGAN, StyleGAN-2, StyleGAN-XL. Our findings suggest that certain visual concepts and structures are inherently embedded in the natural world and can be learned by different models regardless of the specific task or architecture, and without the use of semantic labels. We can visualize shared concepts directly due to ge nerative models included in our analysis. The Rosetta Neurons facilitate model-t o-model translation enabling various inversion-based manipulations, including cr oss-class alignments, shifting, zooming, and more, without the need for speciali zed training.
****
Semi-Supervised Semantic Segmentation under Label Noise via Diverse Learning Gro ups
Peixia Li, Pulak Purkait, Thalaiyasingam Ajanthan, Majid Abdolshah, Ravi Garg, H isham Husain, Chenchen Xu, Stephen Gould, Wanli Ouyang, Anton van den Hengel; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 23, pp. 1229-1238
Semi-supervised semantic segmentation methods use a small amount of clean pixel-level annotations to guide the interpretation of a larger quantity of unlabelled image data. The challenges of providing pixel-accurate annotations at scale mea n that the labels are typically noisy, and this contaminates the final results. In this work, we propose an approach that is robust to label noise in the annota ted data.
The method uses two diverse learning groups with different network architecture s to effectively handle both label noise and unlabelled images. Each learning gr oup consists of a teacher network, a student network and a novel filter module. The filter module of each learning group utilizes pixel-level features from the teacher network to detect incorrectly labelled pixels. To reduce confirmation bi as, we employ the labels cleaned by the filter module from one learning group to train the other learning group. Experimental results on two different benchmark s and settings demonstrate the superiority of our method over state-of-the-art a pproaches.
****
AdaMV-MoE: Adaptive Multi-Task Vision Mixture-of-Experts
Tianlong Chen, Xuxi Chen, Xianzhi Du, Abdullah Rashwan, Fan Yang, Huizhong Chen, Zhangyang Wang, Yeqing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17346-17357
Sparsely activated Mixture-of-Experts (MoE) is becoming a promising paradigm for multi-task learning (MTL). Instead of compressing multiple tasks' knowledge int o a single model, MoE separates the parameter space and only utilizes the releva nt model pieces given task type and its input, which provides stabilized MTL tra ining and ultra-efficient inference. However, current MoE approaches adopt a fix ed network capacity (e.g., two experts in usual) for all tasks. It potentially r esults in the over-fitting of simple tasks or the under-fitting of challenging s

cenarios, especially when tasks are significantly distinctive in their complexity. In this paper, we propose an adaptive MoE framework for multi-task vision recognition, dubbed AdaMV-MoE. Based on the training dynamics, it automatically determines the number of activated experts for each task, avoiding the laborious manual tuning of optimal model size. To validate our proposal, we benchmark it on ImageNet classification and COCO object detection & instance segmentation which are notoriously difficult to learn in concert, due to their discrepancy. Extensive experiments across a variety of vision transformers demonstrate a superior performance of AdaMV-MoE, compared to MTL with a shared backbone and the recent state-of-the-art (SoTA) MTL MoE approach. Codes are available online: https://github.com/google-research/google-research/tree/master/moe_mtl.
****

Hierarchical Visual Categories Modeling: A Joint Representation Learning and Density Estimation Framework for Out-of-Distribution Detection
Jinglun Li, Xinyu Zhou, Pinxue Guo, Yixuan Sun, Yiwen Huang, Weifeng Ge, Wenqiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23425-23435
Detecting out-of-distribution inputs for visual recognition models has become critical in safe deep learning. This paper proposes a novel hierarchical visual category modeling scheme to separate out-of-distribution data from in-distribution data through joint representation learning and statistical modeling. We learn a mixture of Gaussian models for each in-distribution category. There are many Gaussian mixture models to model different visual categories. With these Gaussian models, we design an in-distribution score function by aggregating multiple Mahalanobis-based metrics. We don't use any auxiliary outlier data as training samples, which may hurt the generalization ability of out-of-distribution detection algorithms. We split the ImageNet1k dataset into ten folds randomly. We use one fold as the in-distribution dataset and the others as out-of-distribution datasets to evaluate the proposed method. We also conduct experiments on seven popular benchmarks, including CIFAR, iNaturalist, SUN, Places, Textures, ImageNet-O, and OpenImage-O. Extensive experiments indicate that the proposed method outperforms state-of-the-art algorithms clearly. Meanwhile, we find that our visual representation has a competitive performance when compared with features learned by classical methods. These results demonstrate that the proposed method hasn't weakened the discriminative ability of visual recognition models and keeps high efficiency in detecting out-of-distribution samples.
****

Diffuse3D: Wide-Angle 3D Photography via Bilateral Diffusion
Yutao Jiang, Yang Zhou, Yuan Liang, Wenxi Liu, Jianbo Jiao, Yuhui Quan, Shengfeng He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8998-9008
This paper aims to resolve the challenging problem of wide-angle novel view synthesis from a single image, a.k.a. wide-angle 3D photography. Existing approaches rely on local context and treat them equally to inpaint occluded RGB and depth regions, which fail to deal with large-region occlusion (i.e., observing from an extreme angle) and foreground layers might blend into background inpainting. To address the above issues, we propose Diffuse3D which employs a pre-trained diffusion model for global synthesis, while amending the model to activate depth-aware inference. Our key insight is to alter the convolution mechanism in the denoising process. We inject depth information into the denoising convolution operation with bilateral kernels, i.e., a depth kernel and a spatial kernel, to consider layered correlations among pixels. In this way, foreground regions are overlooked in background inpainting and only pixels close in depth are leveraged. On the other hand, we propose a global-local balancing approach to maximize both contextual understandings. Extensive experiments demonstrate that our approach outperforms state-of-the-art methods in novel view synthesis, especially in wide-angle scenarios. More importantly, our method does not require any training and is a plug-and-play module that can be integrated with any diffusion model. Our code can be found at https://github.com/yutaojiang1/Diffuse3D.
****

ReNeRF: Relightable Neural Radiance Fields with Nearfield Lighting
Yingyan Xu, Gaspard Zoss, Prashanth Chandran, Markus Gross, Derek Bradley, Paulo Gotardo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22581-22591
Recent work on radiance fields and volumetric inverse rendering (e.g., NeRFs) has provided excellent results in building data-driven models of real scenes for novel view synthesis with high photorealism. While full control over viewpoint is achieved, scene lighting is typically "baked" into the model and cannot be changed; other methods only capture limited variation in lighting or make restrictive assumptions about the captured scene. These limitations prevent the application on arbitrary materials and novel 3D environments with complex, distinct lighting. In this paper, we target the application scenario of capturing high-fidelity assets for neural relighting in controlled studio conditions, but without requiring a dense light stage. Instead, we leverage a small number of area lights commonly used in photogrammetry. We propose ReNeRF, a relightable radiance field model based on the intuitive and powerful approach of image-based relighting, which implicitly captures global light transport (for arbitrary objects) without complex, error-prone simulations. Thus, our new method is simple and provides full control over viewpoint and lighting, without simplistic assumptions about how light interacts with the scene. In addition, ReNeRF does not rely on the usual assumption of distant lighting - during training, we explicitly account for the distance between 3D points in the volume and point samples on the light sources. Thus, at test time, we achieve better generalization to novel, continuous lighting directions, including nearfield lighting effects.
****
Segment Anything
Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, Ross Girshick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4015-4026
We introduce the Segment Anything (SA) project: a new task, model, and dataset for image segmentation. Using our efficient model in a data collection loop, we built the largest segmentation dataset to date (by far), with over 1 billion masks on 11M licensed and privacy respecting images. The model is designed and trained to be promptable, so it can transfer zero-shot to new image distributions and tasks. We evaluate its capabilities on numerous tasks and find that its zero-shot performance is impressive -- often competitive with or even superior to prior fully supervised results. We are releasing the Segment Anything Model (SAM) and corresponding dataset (SA-1B) of 1B masks and 11M images at https://segment-anything.com to foster research into foundation models for computer vision. We recommend reading the full paper at: https://arxiv.org/abs/2304.02643.
****
Unsupervised Prompt Tuning for Text-Driven Object Detection
Weizhen He, Weijie Chen, Binbin Chen, Shicai Yang, Di Xie, Luojun Lin, Donglian Qi, Yueting Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2651-2661
Grounded language-image pre-trained models have shown strong zero-shot generalization to various downstream object detection tasks. Despite their promising performance, the models rely heavily on the laborious prompt engineering. Existing works typically address this problem by tuning text prompts using downstream training data in a few-shot or fully supervised manner. However, a rarely studied problem is to optimize text prompts without using any annotations. In this paper, we delve into this problem and propose an Unsupervised Prompt Tuning framework for text-driven object detection, which is composed of two novel mean teaching mechanisms. In conventional mean teaching, the quality of pseudo boxes is expected to optimize better as the training goes on, but there is still a risk of overfitting noisy pseudo boxes. To mitigate this problem, 1) we propose Nested Mean Teaching, which adopts nested-annotation to supervise teacher-student mutual learning in a bi-level optimization manner; 2) we propose Dual Complementary Teaching, which employs an offline pre-trained teacher and an online mean teacher via da

ta-augmentation-based complementary labeling so as to ensure learning without accumulating confirmation bias. By integrating these two mechanisms, the proposed unsupervised prompt tuning framework achieves significant performance improvement on extensive object detection datasets.
****
Tubelet-Contrastive Self-Supervision for Video-Efficient Generalization
Fida Mohammad Thoker, Hazel Doughty, Cees G. M. Snoek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13812-13823
We propose a self-supervised method for learning motion-focused video representations. Existing approaches minimize distances between temporally augmented videos, which maintain high spatial similarity. We instead propose to learn similarities between videos with identical local motion dynamics but an otherwise different appearance. We do so by adding synthetic motion trajectories to videos which we refer to as tubelets.
By simulating different tubelet motions and applying transformations, such as scaling and rotation, we introduce motion patterns beyond what is present in the pretraining data. This allows us to learn a video representation that is remarkably data efficient: our approach maintains performance when using only 25% of the pretraining videos. Experiments on 10 diverse downstream settings demonstrate our competitive performance and generalizability to new domains and fine-grained actions.
****
Re-ReND: Real-Time Rendering of NeRFs across Devices
Sara Rojas, Jesus Zarzar, Juan C. Pérez, Artsiom Sanakoyeu, Ali Thabet, Albert Pumarola, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3632-3641
This paper proposes a novel approach for rendering a pre-trained Neural Radiance Field (NeRF) in real-time on resource-constrained devices. We introduce Re-ReND, a method enabling Real-time Rendering of NeRFs across Devices. Re-ReND is designed to achieve real-time performance by converting the NeRF into a representation that can be efficiently processed by standard graphics pipelines. The proposed method distills the NeRF by extracting the learned density into a mesh, while the learned color information is factorized into a set of matrices that represent the scene's light field. Factorization implies the field is queried via inexpensive MLP-free matrix multiplications, while using a light field allows rendering a pixel by querying the field a single time--as opposed to hundreds of queries when employing a radiance field. Since the proposed representation can be implemented using a fragment shader, it can be directly integrated with standard rasterization frameworks. Our flexible implementation can render a NeRF in real-time with low memory requirements and on a wide range of resource-constrained devices, including mobiles and AR/VR headsets. Notably, we find that Re-ReND can achieve over a 2.6-fold increase in rendering speed versus the state-of-the-art without perceptible losses in quality.
****
360VOT: A New Benchmark Dataset for Omnidirectional Visual Object Tracking
Huajian Huang, Yinzhe Xu, Yingshu Chen, Sai-Kit Yeung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20566-20576
360deg images can provide an omnidirectional field of view which is important for stable and long-term scene perception. In this paper, we explore 360deg images for visual object tracking and perceive new challenges caused by large distortion, stitching artifacts, and other unique attributes of 360deg images. To alleviate these problems, we take advantage of novel representations of target localization, i.e., bounding field-of-view, and then introduce a general 360 tracking framework that can adopt typical trackers for omnidirectional tracking. More importantly, we propose a new large-scale omnidirectional tracking benchmark dataset, 360VOT, in order to facilitate future research. 360VOT contains 120 sequences with up to 113K high-resolution frames in equirectangular projection. And the tracking targets cover 32 categories in diverse scenarios. Moreover, we provide 4 types of unbiased ground truth, including (rotated) bounding boxes and (rotated) bounding field-of-views, as well as new metrics tailored for 360deg images whic

h allow accurate evaluation of omnidirectional tracking performance. Finally, we extensively evaluated 20 state-of-the-art visual trackers and provided a new baseline for future comparisons. Homepage: https://360vot.hkustvgd.com
****
Is Imitation All You Need? Generalized Decision-Making with Dual-Phase Training
Yao Wei, Yanchao Sun, Ruijie Zheng, Sai Vemprala, Rogerio Bonatti, Shuhang Chen, Ratnesh Madaan, Zhongjie Ba, Ashish Kapoor, Shuang Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16221-16231
We introduce DualMind, a generalist agent designed to tackle various decision-making tasks that addresses challenges posed by current methods, such as overfitting behaviors and dependence on task-specific fine-tuning. DualMind uses a novel "Dual-phase" training strategy that emulates how humans learn to act in the world. The model first learns fundamental common knowledge through a self-supervised objective tailored for control tasks and then learns how to make decisions based on different contexts through imitating behaviors conditioned on given prompts. DualMind can handle tasks across domains, scenes, and embodiments using just a single set of model weights and can execute zero-shot prompting without requiring task-specific fine-tuning. We evaluate DualMind on MetaWorld and Habitat through extensive experiments and demonstrate its superior generalizability compared to previous techniques, outperforming other generalist agents by over 50% and 70% on Habitat and MetaWorld, respectively. On the 45 tasks in MetaWorld, DualMind achieves over 30 tasks at a 90% success rate. Our source code is available at https://github.com/yunyikristy/DualMind.
****
Generalizing Event-Based Motion Deblurring in Real-World Scenarios
Xiang Zhang, Lei Yu, Wen Yang, Jianzhuang Liu, Gui-Song Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10734-10744
Event-based motion deblurring has shown promising results by exploiting low-latency events. However, current approaches are limited in their practical usage, as they assume the same spatial resolution of inputs and specific blurriness distributions. This work addresses these limitations and aims to generalize the performance of event-based deblurring in real-world scenarios. We propose a scale-aware network that allows flexible input spatial scales and enables learning from different temporal scales of motion blur. A two-stage self-supervised learning scheme is then developed to fit real-world data distribution. By utilizing the relativity of blurriness, our approach efficiently ensures the restored brightness and structure of latent images and further generalizes deblurring performance to handle varying spatial and temporal scales of motion blur in a self-distillation manner. Our method is extensively evaluated, demonstrating remarkable performance, and we also introduce a real-world dataset consisting of multi-scale blurry frames and events to facilitate research in event-based deblurring.
****
Handwritten and Printed Text Segmentation: A Signature Case Study
Sina Gholamian, Ali Vahdat; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 582-592
While analyzing scanned documents, handwritten text can overlap with printed text. This overlap causes difficulties during the optical character recognition (OCR) and digitization process of documents, and subsequently, hurts downstream NLP tasks. Prior research either focuses solely on the binary classification of handwritten text or performs a three-class segmentation of the document, i.e., recognition of handwritten, printed, and background pixels. This approach results in the assignment of overlapping handwritten and printed pixels to only one of the classes, and thus, they are not accounted for in the other class. Thus, in this research, we develop novel approaches to address the challenges of handwritten and printed text segmentation. Our objective is to recover text from different classes in their entirety, especially enhancing the segmentation performance on overlapping sections. To support this task, we introduce a new dataset, SignaTR6K, collected from real legal documents, as well as a new model architecture for the handwritten and printed text segmentation task. Our best configuration outper

forms prior work on two different datasets by 17.9% and 7.3% on IoU scores. The SignaTR6K dataset is accessible for download via the following link: https://forms.office.com/r/2a5RDg7cAY.
****
LERF: Language Embedded Radiance Fields
Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, Matthew Tancik; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19729-19739
Humans describe the physical world using natural language to refer to specific 3D locations based on a vast range of properties: visual appearance, semantics, abstract associations, or actionable affordances. In this work we propose Language Embedded Radiance Fields (LERFs), a method for grounding language embeddings from off-the-shelf models like CLIP into NeRF, which enable these types of open-ended language queries in 3D. LERF learns a dense, multi-scale language field inside NeRF by volume rendering CLIP embeddings along training rays, supervising these embeddings across training views to provide multi-view consistency and smooth the underlying language field. After optimization, LERF can extract 3D relevancy maps for a broad range of language prompts interactively in real-time, which has potential use cases in robotics, understanding vision-language models, and interacting with 3D scenes. LERF enables pixel-aligned, zero-shot queries on the distilled 3D CLIP embeddings without relying on region proposals or masks, supporting long-tail open-vocabulary queries hierarchically across the volume. See the project website at: https://lerf.io
****
DomainAdaptor: A Novel Approach to Test-time Adaptation
Jian Zhang, Lei Qi, Yinghuan Shi, Yang Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18971-18981
To deal with the domain shift between training and test samples, current methods have primarily focused on learning generalizable features during training and ignore the specificity of unseen samples that are also critical during the test. In this paper, we investigate a more challenging task that aims to adapt a trained CNN model to unseen domains during the test. To maximumly mine the information in the test data, we propose a unified method called DomainAdaptor for the test-time adaptation, which consists of an AdaMixBN module and a Generalized Entropy Minimization (GEM) loss. Specifically, AdaMixBN addresses the domain shift by adaptively fusing training and test statistics in the normalization layer via a dynamic mixture coefficient and a statistic transformation operation. To further enhance the adaptation ability of AdaMixBN, we design a GEM loss that extends the Entropy Minimization loss to better exploit the information in the test data. Extensive experiments show DomainAdaptor consistently outperforms the state-of-the-art methods on four benchmarks. Furthermore, our method brings more remarkable improvement against existing methods on the few-data unseen domain. The code is available at https://github.com/koncle/DomainAdaptor.
****
RCA-NOC: Relative Contrastive Alignment for Novel Object Captioning
Jiashuo Fan, Yaoyuan Liang, Leyao Liu, Shaolun Huang, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15510-15520
In this paper, we introduce a novel approach to novel object captioning which employs relative contrastive learning to learn visual and semantic alignment. Our approach maximizes compatibility between regions and object tags in a contrastive manner. To set up a proper contrastive learning objective, for each image, we augment tags by leveraging the relative nature of positive and negative pairs obtained from foundation models such as CLIP. We then use the rank of each augmented tag in a list as a relative relevance label to contrast each top-ranked tag with a set of lower-ranked tags. This learning objective encourages the top-ranked tags to be more compatible with their image and text context than lower-ranked tags, thus improving the discriminative ability of the learned multi-modality representation. We evaluate our approach on two datasets and show that our proposed RCA-NOC approach outperforms state-of-the-art methods by a large margin, demo

nstrating its effectiveness in improving vision-language representation for novel object captioning.
****
Mitigating and Evaluating Static Bias of Action Representations in the Background and the Foreground
Haoxin Li, Yuan Liu, Hanwang Zhang, Boyang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19911-19923
In video action recognition, shortcut static features can interfere with the learning of motion features, resulting in poor out-of-distribution (OOD) generalization. The video background is clearly a source of static bias, but the video foreground, such as the clothing of the actor, can also provide static bias. In this paper, we empirically verify the existence of foreground static bias by creating test videos with conflicting signals from the static and moving portions of the video. To tackle this issue, we propose a simple yet effective technique, StillMix, to learn robust action representations. Specifically, StillMix identifies bias-inducing video frames using a 2D reference network and mixes them with videos for training, serving as effective bias suppression even when we cannot explicitly extract the source of bias within each video frame or enumerate types of bias. Finally, to precisely evaluate static bias, we synthesize two new benchmarks, SCUBA for static cues in the background, and SCUFO for static cues in the foreground. With extensive experiments, we demonstrate that StillMix mitigates both types of static bias and improves video representations for downstream applications.
****
RbA: Segmenting Unknown Regions Rejected by All
Nazir Nayal, Misra Yavuz, João F. Henriques, Fatma Güney; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 711-722
Standard semantic segmentation models owe their success to curated datasets with a fixed set of semantic categories, without contemplating the possibility of identifying unknown objects from novel categories. Existing methods in outlier detection suffer from a lack of smoothness and objectness in their predictions, due to limitations of the per-pixel classification paradigm. Furthermore, additional training for detecting outliers harms the performance of known classes. In this paper, we explore another paradigm with region-level classification to better segment unknown objects. We show that the object queries in mask classification tend to behave like one vs. all classifiers. Based on this finding, we propose a novel outlier scoring function called RbA by defining the event of being an outlier as being rejected by all known classes. Our extensive experiments show that mask classification improves the performance of the existing outlier detection methods, and the best results are achieved with the proposed RbA. We also propose an objective to optimize RbA using minimal outlier supervision. Further fine-tuning with outliers improves the unknown performance, and unlike previous methods, it does not degrade the inlier performance.
Project page: https://kuis-ai.github.io/RbA
****
CuNeRF: Cube-Based Neural Radiance Field for Zero-Shot Medical Image Arbitrary-Scale Super Resolution
Zixuan Chen, Lingxiao Yang, Jian-Huang Lai, Xiaohua Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21185-21195
Medical image arbitrary-scale super-resolution (MIASSR) has recently gained widespread attention, aiming to supersample medical volumes at arbitrary scales via a single model. However, existing MIASSR methods face two major limitations: (i) reliance on high-resolution (HR) volumes and (ii) limited generalization ability, which restricts their applications in various scenarios. To overcome these limitations, we propose Cube-based Neural Radiance Field (CuNeRF), a zero-shot MIASSR framework that is able to yield medical images at arbitrary scales and free viewpoints in a continuous domain. Unlike existing MISR methods that only fit the mapping between low-resolution (LR) and HR volumes, CuNeRF focuses on building a continuous volumetric representation from each LR volume without the knowledge from the corresponding HR one. This is achieved by the proposed differentiable

modules: cube-based sampling, isotropic volume rendering, and cube-based hierarchical rendering. Through extensive experiments on magnetic resource imaging (MRI) and computed tomography (CT) modalities, we demonstrate that CuNeRF can synthesize high-quality SR medical images, which outperforms state-of-the-art MISR methods, achieving better visual verisimilitude and fewer objectionable artifacts. Compared to existing MISR methods, our CuNeRF is more applicable in practice.
****

Beyond Object Recognition: A New Benchmark towards Object Concept Learning
Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20029-20040
Understanding objects is a central building block of AI, especially for embodied AI. Even though object recognition excels with deep learning, current machines struggle to learn higher-level knowledge, e.g., what attributes an object has, and what we can do with it. Here, we propose a challenging Object Concept Learning (OCL) task to push the envelope of object understanding. It requires machines to reason out affordances and simultaneously give the reason: what attributes make an object possess these affordances. To support OCL, we build a densely annotated knowledge base including extensive annotations for three levels of object concept (category, attribute, affordance), and the clear causal relations of three levels. By analyzing the causal structure of OCL, we present a baseline, Object Concept Reasoning Network (OCRN). It leverages concept instantiation and causal intervention to infer the three levels. In experiments, OCRN effectively infers the object knowledge while following the causalities well. Our data and code are available at https://mvig-rhos.com/ocl.
****

Towards Open-Vocabulary Video Instance Segmentation
Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, Efstratios Gavves; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4057-4066
Video Instance Segmentation (VIS) aims at segmenting and categorizing objects in videos from a closed set of training categories, lacking the generalization ability to handle novel categories in real-world videos. To address this limitation, we make the following three contributions. First, we introduce the novel task of Open-Vocabulary Video Instance Segmentation, which aims to simultaneously segment, track, and classify objects in videos from open-set categories, including novel categories unseen during training. Second, to benchmark Open-Vocabulary VIS, we collect a Large-Vocabulary Video Instance Segmentation dataset (LV-VIS), that contains well-annotated objects from 1,196 diverse categories, significantly surpassing the category size of existing datasets by more than one order of magnitude. Third, we propose an efficient Memory-Induced Transformer architecture, OV2Seg, to first achieve Open-Vocabulary VIS in an end-to-end manner with near real-time inference speed. Extensive experiments on LV-VIS and four existing VIS datasets demonstrate the strong zero-shot generalization ability of OV2Seg on novel categories.
The dataset and code are released here https://github.com/haochenheheda/LVVIS.
****

Unleashing the Power of Gradient Signal-to-Noise Ratio for Zero-Shot NAS
Zihao Sun, Yu Sun, Longxing Yang, Shun Lu, Jilin Mei, Wenxiao Zhao, Yu Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5763-5773
Neural Architecture Search (NAS) aims to automatically find optimal neural network architectures in an efficient way. Zero-Shot NAS is a promising technique that leverages proxies to predict the accuracy of candidate architectures without any training. However, we have observed that most existing proxies do not consistently perform well across different search spaces, and are less concerned with generalization. Recently, the gradient signal-to-noise ratio (GSNR) was shown to be correlated with neural network generalization performance. In this paper, we not only explicitly give the probability that larger GSNR at network initialization can ensure better generalization, but also theoretically prove that GSNR can

ensure better convergence. Then we design the Xi-based gradient signal-to-noise ratio (Xi-GSNR) as a Zero-Shot NAS proxy to predict the network accuracy at initialization. Extensive experiments in different search spaces demonstrate that Xi-GSNR provides superior ranking consistency compared to previous proxies. Moreover, Xi-GSNR-based Zero-Shot NAS also achieves outstanding performance when directly searching for the optimal architecture in various search spaces and datasets. The source code is available at https://github.com/Sunzh1996/Xi-GSNR.
****
EgoObjects: A Large-Scale Egocentric Dataset for Fine-Grained Object Understanding
Chenchen Zhu, Fanyi Xiao, Andres Alvarado, Yasmine Babaei, Jiabo Hu, Hichem El-Mohri, Sean Culatana, Roshan Sumbaly, Zhicheng Yan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20110-20120
Object understanding in egocentric visual data is arguably a fundamental research topic in egocentric vision. However, existing object datasets are either non-egocentric or have limitations in object categories, visual content, and annotation granularities. In this work, we introduce EgoObjects, a large-scale egocentric dataset for fine-grained object understanding. Its Pilot version contains over 9K videos collected by 250 participants from 50+ countries using 4 wearable devices, and over 650K object annotations from 368 object categories. Unlike prior datasets containing only object category labels, EgoObjects also annotates each object with an instance-level identifier, and includes over 14K unique object instances. EgoObjects was designed to capture the same object under diverse background complexities, surrounding objects, distance, lighting and camera motion. In parallel to the data collection, we conducted data annotation by developing a multi-stage federated annotation process to accommodate the growing nature of the dataset. To bootstrap the research on EgoObjects, we present a suite of 4 benchmark tasks around the egocentric object understanding, including a novel instance level- and the classical category level object detection. Moreover, we also introduce 2 novel continual learning object detection tasks. The dataset and API are available at https://github.com/facebookresearch/EgoObjects.
****
What Can Simple Arithmetic Operations Do for Temporal Modeling?
Wenhao Wu, Yuxin Song, Zhun Sun, Jingdong Wang, Chang Xu, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13712-13722
Temporal modeling plays a crucial role in understanding video content. To tackle this problem, previous studies built complicated temporal relations through time sequence thanks to the development of computationally powerful devices. In this work, we explore the potential of four simple arithmetic operations for temporal modeling. Specifically, we first capture auxiliary temporal cues by computing addition, subtraction, multiplication, and division between pairs of extracted frame features. Then, we extract corresponding features from these cues to benefit the original temporal-irrespective domain. We term such a simple pipeline as an Arithmetic Temporal Module (ATM), which operates on the stem of a visual backbone with a plug-and-play style. We conduct comprehensive ablation studies on the instantiation of ATMs and demonstrate that this module provides powerful temporal modeling capability at a low computational cost. Moreover, the ATM is compatible with both CNNs- and ViTs-based architectures. Our results show that ATM achieves superior performance over several popular video benchmarks. Specifically, on Something-Something V1, V2 and Kinetics-400, we reach top-1 accuracy of 65.6%, 74.6%, and 89.4% respectively. The code is available at https://github.com/whwu95/ATM.
****
Pixel Adaptive Deep Unfolding Transformer for Hyperspectral Image Reconstruction
Miaoyu Li, Ying Fu, Ji Liu, Yulun Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12959-12968
Hyperspectral Image (HSI) reconstruction has made gratifying progress with the deep unfolding framework by formulating the problem into a data module and a prior module. Nevertheless, existing methods still face the problem of insufficient

matching with HSI data. The issues lie in three aspects: 1) fixed gradient descent step in the data module while the degradation of HSI is agnostic in the pixel-level. 2) inadequate prior module for 3D HSI cube. 3) stage interaction ignoring the differences in features at different stages. To address these issues, in this work, we propose a Pixel Adaptive Deep Unfolding Transformer (PADUT) for HSI reconstruction. In the data module, a pixel adaptive descent step is employed to focus on pixel-level agnostic degradation. In the prior module, we introduce the Non-local Spectral Transformer (NST) to emphasize the 3D characteristics of HSI for recovering. Moreover, inspired by the diverse expression of features in different stages and depths, the stage interaction is improved by the Fast Fourier Transform (FFT). Experimental results on both simulated and real scenes exhibit the superior performance of our method compared to state-of-the-art HSI reconstruction methods. The code is released at: https://github.com/MyuLi/PADUT
****
BiViT: Extremely Compressed Binary Vision Transformers
Yefei He, Zhenyu Lou, Luoming Zhang, Jing Liu, Weijia Wu, Hong Zhou, Bohan Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5651-5663
Model binarization can significantly compress model size, reduce energy consumption, and accelerate inference through efficient bit-wise operations. Although binarizing convolutional neural networks have been extensively studied, there is little work on exploring binarization of vision Transformers which underpin most recent breakthroughs in visual recognition. To this end, we propose to solve two fundamental challenges to push the horizon of Binary Vision Transformers (BiViT). First, the traditional binary method does not take the long-tailed distribution of softmax attention into consideration, bringing large binarization errors in the attention module. To solve this, we propose Softmax-aware Binarization, which dynamically adapts to the data distribution and reduces the error caused by binarization. Second, to better preserve the information of the pretrained model and restore accuracy, we propose a Cross-layer Binarization scheme that decouples the binarization of self-attention and multi-layer perceptrons (MLPs), and Parameterized Weight Scales which introduce learnable scaling factors for weight binarization. Overall, our method performs favorably against state-of-the-arts by 19.8% on the TinyImageNet dataset. On ImageNet, our BiViT achieves a competitive 75.6% Top-1 accuracy over Swin-S model. Additionally, on COCO object detection, our method achieves an mAP of 40.8 with a Swin-T backbone over Cascade Mask R-CNN framework.
****
Dynamic PlenOctree for Adaptive Sampling Refinement in Explicit NeRF
Haotian Bai, Yiqi Lin, Yize Chen, Lin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8785-8795
The explicit neural radiance field (NeRF) has gained considerable interest for its efficient training and fast inference capabilities, making it a promising direction such as virtual reality and gaming. In particular, PlenOctree (POT), an explicit hierarchical multi-scale octree representation, has emerged as a structural and influential framework. However, POT's fixed structure for direct optimization is sub-optimal as the scene complexity evolves continuously with updates to cached color and density, necessitating refining the sampling distribution to capture signal complexity accordingly. To address this issue, we propose the dynamic PlenOctree (DOT), which adaptively refines the sample distribution to adjust to changing scene complexity. Specifically, DOT proposes a concise yet novel hierarchical feature fusion strategy during the iterative rendering process. Firstly, it identifies the regions of interest through training signals to ensure adaptive and efficient refinement. Next, rather than directly filtering out valueless nodes, DOT introduces the sampling and pruning operations for octrees to aggregate features, enabling rapid parameter learning. Compared with POT, our DOT outperforms it by enhancing visual quality, reducing over 55.15/68.84% parameters, and providing 1.7/1.9 times FPS for NeRF-synthetic and Tanks & Temples, respectively.
****

Scene Matters: Model-based Deep Video Compression
Lv Tang, Xinfeng Zhang, Gai Zhang, Xiaoqi Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12481-12491
Video compression has always been a popular research area, where many traditional and deep video compression methods have been proposed. These methods typically rely on signal prediction theory to enhance compression performance by designing high efficient intra and inter prediction strategies and compressing video frames one by one. In this paper, we propose a novel model-based video compression (MVC) framework that regards scenes as the fundamental units for video sequences. Our proposed MVC directly models the intensity variation of the entire video sequence in one scene, seeking non-redundant representations instead of reducing redundancy through spatio-temporal predictions. To achieve this, we employ implicit neural representation as our basic modeling architecture. To improve the efficiency of video modeling, we first propose context-related spatial positional embedding and frequency domain supervision in spatial context enhancement. For temporal correlation capturing, we design the scene flow constrain mechanism and temporal contrastive loss. Extensive experimental results demonstrate that our method achieves up to a 20% bitrate reduction compared to the latest video coding standard H.266 and is more efficient in decoding than existing video coding strategies.
****
Tree-Structured Shading Decomposition
Chen Geng, Hong-Xing Yu, Sharon Zhang, Maneesh Agrawala, Jiajun Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 488-498
We study inferring a tree-structured representation from a single image for object shading. Prior work typically uses the parametric or measured representation to model shading, which is neither interpretable nor easily editable. We propose using the shade tree representation, which combines basic shading nodes and compositing methods to factorize object surface shading. The shade tree representation enables novice users who are unfamiliar with the physical shading process to edit object shading in an efficient and intuitive manner. A main challenge in inferring the shade tree is that the inference problem involves both the discrete tree structure and the continuous parameters of the tree nodes. We propose a hybrid approach to address this issue. We introduce an auto-regressive inference model to generate a rough estimation of the tree structure and node parameters, and then we fine-tune the inferred shade tree through an optimization algorithm. We show experiments on synthetic images, captured reflectance, real images, and non-realistic vector drawings, allowing downstream applications such as material editing, vectorized shading, and relighting. Project website: https://chen-geng.com/inv-shade-trees.
****
EfficientTrain: Exploring Generalized Curriculum Learning for Training Visual Backbones
Yulin Wang, Yang Yue, Rui Lu, Tianjiao Liu, Zhao Zhong, Shiji Song, Gao Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5852-5864
The superior performance of modern deep networks usually comes with a costly training procedure. This paper presents a new curriculum learning approach for the efficient training of visual backbones (e.g., vision Transformers). Our work is inspired by the inherent learning dynamics of deep networks: we experimentally show that at an earlier training stage, the model mainly learns to recognize some 'easier-to-learn' discriminative patterns within each example, e.g., the lower-frequency components of images and the original information before data augmentation. Driven by this phenomenon, we propose a curriculum where the model always leverages all the training data at each epoch, while the curriculum starts with only exposing the 'easier-to-learn' patterns of each example, and introduces gradually more difficult patterns. To implement this idea, we 1) introduce a cropping operation in the Fourier spectrum of the inputs, which enables the model to learn from only the lower-frequency components efficiently, 2) demonstrate that e

xposing the features of original images amounts to adopting weaker data augmentation, and 3) integrate 1) and 2) and design a curriculum learning schedule with a greedy-search algorithm. The resulting approach, EfficientTrain, is simple, general, yet surprisingly effective. As an off-the-shelf method, it reduces the wall-time training cost of a wide variety of popular models (e.g., ResNet, ConvNeXt, DeiT, PVT, Swin, and CSWin) by >1.5x on ImageNet-1K/22K without sacrificing accuracy. It is also effective for self-supervised learning (e.g., MAE). Code is available at https://github.com/LeapLabTHU/EfficientTrain.
****
Simulating Fluids in Real-World Still Images
Siming Fan, Jingtan Piao, Chen Qian, Hongsheng Li, Kwan-Yee Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15922-15931
In this work, we tackle the problem of real-world fluid animation from a still image. The key of our system is a surface-based layered representation, where the scene is decoupled into a surface fluid layer and an impervious background layer with corresponding transparencies to characterize the composition of the two layers. The animated video can be produced by warping only the surface fluid layer according to the estimation of fluid motions and recombining it with the background. In addition, we introduce surface-only fluid simulation, a 2.5D fluid calculation, as a replacement for motion estimation.
Specifically, we leverage triangular mesh based on a monocular depth estimator to represent fluid surface layer and simulate the motion with the inspiration of classic physics theory of hybrid Lagrangian-Eulerian method, along with a learnable network so as to adapt to complex real-world image textures.Extensive experiments not only indicate our method's competitive performance for common fluid scenes but also better robustness and reasonability under complex transparent fluid scenarios. Moreover, as proposed surface-based layer representation and surface-only fluid simulation naturally disentangle the scene, interactive editing such as adding objects and texture replacing could be easily achieved with realistic results.
****
SC3K: Self-supervised and Coherent 3D Keypoints Estimation from Rotated, Noisy, and Decimated Point Cloud Data
Mohammad Zohaib, Alessio Del Bue; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22509-22519
This paper proposes a new method to infer keypoints from arbitrary object categories in practical scenarios where point cloud data (PCD) are noisy, down-sampled and arbitrarily rotated. Our proposed model adheres to the following principles: i) keypoints inference is fully unsupervised (no annotation given), ii) keypoints position error should be low and resilient to PCD perturbations (robustness), iii) keypoints should not change their indexes for the intra-class objects (semantic coherence), iv) keypoints should be close to or proximal to PCD surface (compactness). We achieve these desiderata by proposing a new self-supervised training strategy for keypoints estimation that does not assume any a priori knowledge of the object class, and a model architecture with coupled auxiliary losses that promotes the desired keypoints properties. We compare the keypoints estimated by the proposed approach with those of the state-of-the-art unsupervised approaches. The experiments show that our approach outperforms by estimating keypoints with improved coverage (+9.41%) while being semantically consistent (+4.66%) that best characterizes the object's 3D shape for downstream tasks.
****
IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis
Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, Guofeng Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 339-351
Existing inverse rendering combined with neural rendering methods can only perform editable novel view synthesis on object-specific scenes, while we present intrinsic neural radiance fields, dubbed IntrinsicNeRF, which introduce intrinsic d

ecomposition into the NeRF-based neural rendering method and can extend its application to room-scale scenes. Since intrinsic decomposition is a fundamentally under-constrained inverse problem, we propose a novel distance-aware point sampling and adaptive reflectance iterative clustering optimization method, which enables IntrinsicNeRF with traditional intrinsic decomposition constraints to be trained in an unsupervised manner, resulting in multi-view consistent intrinsic decomposition results. To cope with the problem that different adjacent instances of similar reflectance in a scene are incorrectly clustered together, we further propose a hierarchical clustering method with coarse-to-fine optimization to obtain a fast hierarchical indexing representation. It supports compelling real-time augmented applications such as recoloring and illumination variation. Extensive experiments and editing samples on both object-specific/room-scale scenes and synthetic/real-word data demonstrate that we can obtain consistent intrinsic decomposition results and high-fidelity novel view synthesis even for challenging sequences.
****
Segmenting Known Objects and Unseen Unknowns without Prior Knowledge
Stefano Gasperini, Alvaro Marcos-Ramiro, Michael Schmidt, Nassir Navab, Benjamin Busam, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19321-19332
Panoptic segmentation methods assign a known class to each pixel given in input. Even for state-of-the-art approaches, this inevitably enforces decisions that systematically lead to wrong predictions for objects outside the training categories. However, robustness against out-of-distribution samples and corner cases is crucial in safety-critical settings to avoid dangerous consequences. Since real-world datasets cannot contain enough data points to adequately sample the long tail of the underlying distribution, models must be able to deal with unseen and unknown scenarios as well. Previous methods targeted this by re-identifying already-seen unlabeled objects. In this work, we propose the necessary step to extend segmentation with a new setting which we term holistic segmentation. Holistic segmentation aims to identify and separate objects of unseen, unknown categories into instances without any prior knowledge about them while performing panoptic segmentation of known classes. We tackle this new problem with U3HS, which finds unknowns as highly uncertain regions and clusters their corresponding instance-aware embeddings into individual objects. By doing so, for the first time in panoptic segmentation with unknown objects, our U3HS is trained without unknown categories, reducing assumptions and leaving the settings as unconstrained as in real-life scenarios. Extensive experiments on public data from MS COCO, Cityscapes, and Lost&Found demonstrate the effectiveness of U3HS for this new, challenging, and assumptions-free setting called holistic segmentation. Project page: https://holisticseg.github.io.
****
A Good Student is Cooperative and Reliable: CNN-Transformer Collaborative Learning for Semantic Segmentation
Jinjing Zhu, Yunhao Luo, Xu Zheng, Hao Wang, Lin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11720-11730
In this paper, we strive to answer the question 'how to collaboratively learn convolutional neural network (CNN)-based and vision transformer (ViT)-based models by selecting and exchanging the reliable knowledge between them for semantic segmentation?' Accordingly, we propose an online knowledge distillation (KD) framework that can simultaneously learn compact yet effective CNN-based and ViT-based models with two key technical breakthroughs to take full advantage of CNNs and ViT while compensating their limitations. Firstly, we propose heterogeneous feature distillation (HFD) to improve students' consistency in low-layer feature space by mimicking heterogeneous features between CNNs and ViT. Secondly, to facilitate the two students to learn reliable knowledge from each other, we propose bidirectional selective distillation (BSD) that can dynamically transfer selective knowledge. This is achieved by 1) region-wise BSD determining the directions of knowledge transferred between the corresponding regions in the feature space and 2) pixel-wise BSD discerning which of the prediction knowledge to be transferr

ed in the logit space. Extensive experiments on three benchmark datasets demonstrate that our proposed framework outperforms the state-of-the-art online distillation methods by a large margin, and shows its efficacy in learning collaboratively between ViT-based and CNN-based models.
****

CMDA: Cross-Modality Domain Adaptation for Nighttime Semantic Segmentation
Ruihao Xia, Chaoqiang Zhao, Meng Zheng, Ziyan Wu, Qiyu Sun, Yang Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21572-21581
Most nighttime semantic segmentation studies are based on domain adaptation approaches and image input. However, limited by the low dynamic range of conventional cameras, images fail to capture structural details and boundary information in low-light conditions. Event cameras, as a new form of vision sensors, are complementary to conventional cameras with their high dynamic range. To this end, we propose a novel unsupervised Cross-Modality Domain Adaptation (CMDA) framework to leverage multi-modality (Images and Events) information for nighttime semantic segmentation, with only labels on daytime images. In CMDA, we design the Image Motion-Extractor to extract motion information and the Image Content-Extractor to extract content information from images, in order to bridge the gap between different modalities (Images to Events) and domains (Day to Night). Besides, we introduce the first image-event nighttime semantic segmentation dataset. Extensive experiments on both the public image dataset and the proposed image-event dataset demonstrate the effectiveness of our proposed approach. We open-source our code, models, and dataset at https://github.com/XiaRho/CMDA.
****

Learning with Diversity: Self-Expanded Equalization for Better Generalized Deep Metric Learning
Jiexi Yan, Zhihui Yin, Erkun Yang, Yanhua Yang, Heng Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19365-19374
Exploring good generalization ability is essential in deep metric learning (DML). Most existing DML methods focus on improving the model robustness against category shift to keep the performance on unseen categories. However, in addition to category shift, domain shift also widely exists in real-world scenarios. Therefore, learning better generalization ability for the DML model is still a challenging yet realistic problem. In this paper, we propose a new self-expanded equalization (SEE) method to effectively generalize the DML model to both unseen categories and domains. Specifically, we take a `min-max' strategy combined with a proxy-based loss to adaptively augment diverse out-of-distribution samples that vastly expand the span of original training data. To take full advantage of the implicit cross-domain relations between source and augmented samples, we introduce a domain-aware equalization module to induce the domain-invariant distance metric by regularizing the feature distribution in the metric space. Extensive experiments on two benchmarks and a large-scale multi-domain dataset demonstrate the superiority of our SEE over the existing DML methods.
****

Fan-Beam Binarization Difference Projection (FB-BDP): A Novel Local Object Descriptor for Fine-Grained Leaf Image Retrieval
Xin Chen, Bin Wang, Yongsheng Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11102-11111
Fine-grained leaf image retrieval (FGLIR) aims to search similar leaf images in subspecies level which involves very high interclass visual similarity and accordingly poses great challenges to leaf image description. In this study, we introduce a new concept, named fan-beam binarization difference projection (FB-BDP) to address this challenging issue. It is designed based on the theory of fan-beam projection (FBP) which is a mathematical tool originally used for computed tomographic reconstruction of objects and has the merits of capturing the inner structure information of objects in multiple directions and excellent ability to suppress image noise. However, few studies have been made to apply FBP to the description of texture patterns. Rather than calculating ray integrals over the whole

object area, FB-BDP restricts its ray integrals calculated over local patches to guarantee the locality of the extracted features. By binarizing the intensity-differences between the off-center and center rays, FB-BDP enable its ray integrals insensitive to illumination change and more discriminative in the characterization of texture patterns. In additional, due to inheriting the merits of FBP, the proposed FB-BDP is superior over the existing local image descriptors by its invariance to scaling transformation, robustness to noise, and strong ability to capture direction and structure texture patterns. The results of extensive experiments on FGLIR show its higher retrieval accuracy over the benchmark methods, promising generalization power and strong complementarity to deep features.
****
Dynamic Residual Classifier for Class Incremental Learning
Xiuwei Chen, Xiaobin Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18743-18752
The rehearsal strategy is widely used to alleviate the catastrophic forgetting problem in class incremental learning (CIL) by preserving limited exemplars from previous tasks. With imbalanced sample numbers between old and new classes, the classifier learning can be biased. Existing CIL methods exploit the long-tailed (LT) recognition techniques, e.g., the adjusted losses and the data re-sampling methods, to handle the data imbalance issue within each increment task. In this work, the dynamic nature of data imbalance in CIL is shown and a novel Dynamic Residual Classifier (DRC) is proposed to handle this challenging scenario. Specifically, DRC is built upon a recent advance residual classifier with the branch layer merging to handle the model-growing problem. Moreover, DRC is compatible with different CIL pipelines and substantially improves them. Combining DRC with the model adaptation and fusion (MAF) pipeline, this method achieves state-of-the-art results on both the conventional CIL and the LT-CIL benchmarks. Extensive experiments are also conducted for a detailed analysis. The code is publicly available.
****
Optimizing the Placement of Roadside LiDARs for Autonomous Driving
Wentao Jiang, Hao Xiang, Xinyu Cai, Runsheng Xu, Jiaqi Ma, Yikang Li, Gim Hee Lee, Si Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18381-18390
Multi-agent cooperative perception is an increasingly popular topic in the field of autonomous driving, where roadside LiDARs play an essential role. However, how to optimize the placement of roadside LiDARs is a crucial but often overlooked problem. This paper proposes an approach to optimize the placement of roadside LiDARs by selecting optimized positions within the scene for better perception performance. To efficiently obtain the best combination of locations, a greedy algorithm based on the perceptual gain is proposed, which selects the location that can maximize the perceptual gain sequentially. We define perceptual gain as the increased perceptual capability when a new LiDAR is placed. To obtain the perception capability, we propose a perception predictor that learns to evaluate LiDAR placement using only a single point cloud frame. A dataset named Roadside-Opt is created using the CARLA simulator to facilitate research on the roadside LiDAR placement problem. Extensive experiments are conducted to demonstrate the effectiveness of our proposed method.
****
Diverse Inpainting and Editing with GAN Inversion
Ahmet Burak Yildirim, Hamza Pehlivan, Bahri Batuhan Bilecen, Aysegul Dundar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23120-23130
Recent inversion methods have shown that real images can be inverted into StyleGAN's latent space and numerous edits can be achieved on those images thanks to the semantically rich feature representations of well-trained GAN models. However, extensive research has also shown that image inversion is challenging due to the trade-off between high-fidelity reconstruction and editability. In this paper, we tackle an even more difficult task, inverting erased images into GAN's latent space for realistic inpaintings and editings. Furthermore, by augmenting inve

rted latent codes with different latent samples, we achieve diverse inpaintings. Specifically, we propose to learn an encoder and mixing network to combine encoded features from erased images with StyleGAN's mapped features from random samples. To encourage the mixing network to utilize both inputs, we train the networks with generated data via a novel set-up. We also utilize higher-rate features to prevent color inconsistencies between the inpainted and unerased parts. We run extensive experiments and compare our method with state-of-the-art inversion and inpainting methods. Qualitative metrics and visual comparisons show significant improvements.
****
InterDiff: Generating 3D Human-Object Interactions with Physics-Informed Diffusion
Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, Liang-Yan Gui; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14928-14940
This paper addresses a novel task of anticipating 3D human-object interactions (HOIs). Most existing research on HOI synthesis lacks comprehensive whole-body interactions with dynamic objects, e.g., often limited to manipulating small or static objects. Our task is significantly more challenging, as it requires modeling dynamic objects with various shapes, capturing whole-body motion, and ensuring physically valid interactions. To this end, we propose InterDiff, a framework comprising two key steps: (i) interaction diffusion, where we leverage a diffusion model to encode the distribution of future human-object interactions; (ii) interaction correction, where we introduce a physics-informed predictor to correct denoised HOIs in a diffusion step. Our key insight is to inject prior knowledge that the interactions under reference with respect to contact points follow a simple pattern and are easily predictable. Experiments on multiple human-object interaction datasets demonstrate the effectiveness of our method for this task, capable of producing realistic, vivid, and remarkably long-term 3D HOI predictions.
****
DiFaReli: Diffusion Face Relighting
Puntawat Ponglertnapakorn, Nontawat Tritrong, Supasorn Suwajanakorn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22646-22657
We present a novel approach to single-view face relighting in the wild. Handling non-diffuse effects, such as global illumination or cast shadows, has long been a challenge in face relighting. Prior work often assumes Lambertian surfaces, simplified lighting models or involves estimating 3D shape, albedo, or a shadow map. This estimation, however, is error-prone and requires many training examples with lighting ground truth to generalize well. Our work bypasses the need for accurate estimation of intrinsic components and can be trained solely on 2D images without any light stage data, multi-view images, or lighting ground truth. Our key idea is to leverage a conditional diffusion implicit model (DDIM) for decoding a disentangled light encoding along with other encodings related to 3D shape and facial identity inferred from off-the-shelf estimators. We also propose a novel conditioning technique that eases the modeling of the complex interaction between light and geometry by using a rendered shading reference to spatially modulate the DDIM. We achieve state-of-the-art performance on standard benchmark Multi-PIE and can photorealistically relight in-the-wild images. Please visit our page: https://diffusion-face-relighting.github.io.
****
IST-Net: Prior-Free Category-Level Pose Estimation with Implicit Space Transformation
Jianhui Liu, Yukang Chen, Xiaoqing Ye, Xiaojuan Qi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13978-13988
Category-level 6D pose estimation aims to predict the poses and sizes of unseen objects from a specific category. Thanks to prior deformation, which explicitly adapts a category-specific 3D prior (i.e., a 3D template) to a given object instance, prior-based methods attained great success and have become a major research stream. However, obtaining category-specific priors requires collecting a larg

e amount of 3D models, which is labor-consuming and often not accessible in prac tice. This motivates us to investigate whether priors are necessary to make prio r-based methods effective. Our empirical study shows that the 3D prior itself is not the credit to the high performance. The keypoint actually is the explicit d eformation process, which aligns camera and world coordinates supervised by worl d space 3D models (also called canonical space). Inspired by these observations, we introduce a simple prior-free implicit space transformation network, namely IST-Net, to transform camera-space features to world-space counterparts and buil d correspondences between them in an implicit manner without relying on 3D prior s. Besides, we design camera- and world-space enhancers to enrich the features w ith pose-sensitive information and geometrical constraints, respectively. Albeit simple, IST-Net achieves state-of-the-art performance based-on prior-free desig n, with top inference speed on the REAL275 benchmark. Our code and models are av ailable at https://github.com/CVMI-Lab/IST-Net.
****
Building3D: A Urban-Scale Dataset and Benchmarks for Learning Roof Structures fr om Point Clouds
Ruisheng Wang, Shangfeng Huang, Hongxin Yang; Proceedings of the IEEE/CVF Intern ational Conference on Computer Vision (ICCV), 2023, pp. 20076-20086
Urban modeling from LiDAR point clouds is an important topic in computer vision, computer graphics, photogrammetry and remote sensing. 3D city models have found a wide range of applications in smart cities, autonomous navigation, urban plan ning and mapping etc. However, existing datasets for 3D modeling mainly focus on common objects such as furniture or cars. Lack of building datasets has become a major obstacle for applying deep learning technology to specific domains such as urban modeling. In this paper, we present a urban-scale dataset consisting of more than 160 thousands buildings along with corresponding point clouds, mesh a nd wire-frame models, covering 16 cities in Estonia about 998 Km2. We extensivel y evaluate performance of state-of-the-art algorithms including handcrafted and deep feature based methods. Experimental results indicate that Building3D has ch allenges of high intra-class variance, data imbalance and large-scale noises. Th e Building3D is the first and largest urban-scale building modeling benchmark, a llowing a comparison of supervised and self-supervised learning methods. We beli eve that our Building3D will facilitate future research on urban modeling, aeria l path planning, mesh simplification, and semantic/part segmentation etc.
****
Multi-Object Discovery by Low-Dimensional Object Motion
Sadra Safadoust, Fatma Güney; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2023, pp. 734-744
Recent work in unsupervised multi-object segmentation shows impressive results b y predicting motion from a single image despite the inherent ambiguity in predic ting motion without the next image. On the other hand, the set of possible motio ns for an image can be constrained to a low-dimensional space by considering the scene structure and moving objects in it. We propose to model pixel-wise geomet ry and object motion to remove ambiguity in reconstructing flow from a single im age. Specifically, we divide the image into coherently moving regions and use de pth to construct flow bases that best explain the observed flow in each region. We achieve state-of-the-art results in unsupervised multi-object segmentation on synthetic and real-world datasets by modeling the scene structure and object mo tion. Our evaluation of the predicted depth maps shows reliable performance in m onocular depth estimation.
****
Localizing Object-Level Shape Variations with Text-to-Image Diffusion Models
Or Patashnik, Daniel Garibi, Idan Azuri, Hadar Averbuch-Elor, Daniel Cohen-Or; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 023, pp. 23051-23061
Text-to-image models give rise to workflows which often begin with an exploratio n step, where users sift through a large collection of generated images. The glo bal nature of the text-to-image generation process prevents users from narrowing their exploration to a particular object in the image. In this paper, we presen

t a technique to generate a collection of images that depicts variations in the shape of a specific object, enabling an object-level shape exploration process. Creating plausible variations is challenging as it requires control over the shape of the generated object while respecting its semantics. A particular challenge when generating object variations is accurately localizing the manipulation applied over the object's shape. We introduce a prompt-mixing technique that switches between prompts along the denoising process to attain a variety of shape choices. To localize the image-space operation, we present two techniques that use the self-attention layers in conjunction with the cross-attention layers. Moreover, we show that these localization techniques are general and effective beyond the scope of generating object variations. Extensive results and comparisons demonstrate the effectiveness of our method in generating object variations, and the competence of our localization techniques.
****
CoSign: Exploring Co-occurrence Signals in Skeleton-based Continuous Sign Language Recognition
Peiqi Jiao, Yuecong Min, Yanan Li, Xiaotao Wang, Lei Lei, Xilin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20676-20686
The co-occurrence signals (e.g., hand shape, facial expression, and lip pattern) play a critical role in Continuous Sign Language Recognition (CSLR). Compared to RGB data, skeleton data provide a more efficient and concise option, and lay a good foundation for the co-occurrence exploration in CSLR. However, skeleton data are often used as a tool to assist visual grounding and have not attracted sufficient attention. In this paper, we propose a simple yet effective GCN-based approach, named CoSign, to incorporate Co-occurrence Signals and explore the potential of skeleton data in CSLR. Specifically, we propose a group-specific GCN to better exploit the knowledge of each signal and a complementary regularization to prevent complex co-adaptation across signals. Furthermore, we propose a two-stream framework that gradually fuses both static and dynamic information in skeleton data. Experimental results on three public CSLR datasets (PHOENIX14, PHOENIX14-T and CSL-Daily) show that the proposed CoSign achieves competitive performance with recent video-based approaches while reducing the computation cost during training.
****
GACE: Geometry Aware Confidence Enhancement for Black-Box 3D Object Detectors on LiDAR-Data
David Schinagl, Georg Krispel, Christian Fruhwirth-Reisinger, Horst Possegger, Horst Bischof; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6566-6576
Widely-used LiDAR-based 3D object detectors often neglect fundamental geometric information readily available from the object proposals in their confidence estimation. This is mostly due to architectural design choices, which were often adopted from the 2D image domain, where geometric context is rarely available. In 3D, however, considering the object properties and its surroundings in a holistic way is important to distinguish between true and false positive detections, e.g. occluded pedestrians in a group. To address this, we present GACE, an intuitive and highly efficient method to improve the confidence estimation of a given black-box 3D object detector. We aggregate geometric cues of detections and their spatial relationships, which enables us to properly assess their plausibility and consequently, improve the confidence estimation. This leads to consistent performance gains over a variety of state-of-the-art detectors. Across all evaluated detectors, GACE proves to be especially beneficial for the vulnerable road user classes, i.e. pedestrians and cyclists.
****
Curvature-Aware Training for Coordinate Networks
Hemanth Saratchandran, Shin-Fang Chng, Sameera Ramasinghe, Lachlan MacDonald, Simon Lucey; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13328-13338
Coordinate networks are widely used in computer vision due to their ability to r

epresent signals as compressed, continuous entities. However, training these net works with first-order optimizers can be slow, hindering their use in real-time applications. Recent works have opted for shallow voxel-based representations to achieve faster training, but this sacrifices memory efficiency. This work propo ses a solution that leverages second-order optimization methods to significantly reduce training times for coordinate networks while maintaining their compressi bility. Experiments demonstrate the effectiveness of this approach on various si gnal modalities, such as audio, images, videos, shape and neural radiance fields (NeRF).
****
Disentangle then Parse: Night-time Semantic Segmentation with Illumination Disen tanglement
Zhixiang Wei, Lin Chen, Tao Tu, Pengyang Ling, Huaian Chen, Yi Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21 593-21603
Most prior semantic segmentation methods have been developed for day-time scenes , while typically underperforming in night-time scenes due to insufficient and c omplicated lighting conditions. In this work, we tackle this challenge by propos ing a novel night-time semantic segmentation paradigm, i.e., disentangle then pa rse (DTP). DTP explicitly disentangles night-time images into light-invariant re flectance and light-specific illumination components and then recognizes semanti cs based on their adaptive fusion. Concretely, the proposed DTP comprises two ke y components: 1) Instead of processing lighting-entangled features as in prior w orks, our Semantic-Oriented Disentanglement (SOD) framework enables the extracti on of reflectance component without being impeded by lighting, allowing the netw ork to consistently recognize the semantics under cover of varying and complicat ed lighting conditions. 2) Based on the observation that the illumination compon ent can serve as a cue for some semantically confused regions, we further introd uce an Illumination-Aware Parser (IAParser) to explicitly learn the correlation between semantics and lighting, and aggregate the illumination features to yield more precise predictions. Extensive experiments on the night-time segmentation task with various settings demonstrate that DTP significantly outperforms state-of-the-art methods. Furthermore, with negligible additional parameters, DTP can be directly used to benefit existing day-time methods for night-time segmentatio n. Code and dataset are available at https://github.com/w1oves/DTP.git.
****
Large-Scale Land Cover Mapping with Fine-Grained Classes via Class-Aware Semi-Su pervised Semantic Segmentation
Runmin Dong, Lichao Mou, Mengxuan Chen, Weijia Li, Xin-Yi Tong, Shuai Yuan, Lixi an Zhang, Juepeng Zheng, Xiaoxiang Zhu, Haohuan Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16783-16793
Semi-supervised learning has attracted increasing attention in the large-scale l and cover mapping task. However, existing methods overlook the potential to alle viate the class imbalance problem by selecting a suitable set of unlabeled data. Besides, in class-imbalanced scenarios, existing pseudo-labeling methods mostly only pick confident samples, failing to exploit the hard samples during trainin g. To tackle these issues, we propose a unified Class-Aware Semi-Supervised Sema ntic Segmentation framework. The proposed framework consists of three key compon ents. To construct a better semi-supervised learning dataset, we propose a class -aware unlabeled data selection method that is more balanced towards the minorit y classes. Based on the built dataset with improved class balance, we propose a Class-Balanced Cross Entropy loss, jointly considering the annotation bias and t he class bias to re-weight the loss in both sample and class levels to alleviate the class imbalance problem. Moreover, we propose the Class Center Contrast met hod to jointly utilize the labeled and unlabeled data. Specifically, we decompos e the feature embedding space using the ground truth and pseudo-labels, and empl oy the embedding centers for hard and easy samples of each class per image in th e contrast loss to exploit the hard samples during training. Compared with state -of-the-art class-balanced pseudo-labeling methods, the proposed method improves the mean accuracy and mIoU by 4.28% and 1.70%, respectively, on the large-scale

Sentinel-2 dataset with 24 land cover classes.
****
ToonTalker: Cross-Domain Face Reenactment
Yuan Gong, Yong Zhang, Xiaodong Cun, Fei Yin, Yanbo Fan, Xuan Wang, Baoyuan Wu, Yujiu Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7690-7700
We target cross-domain face reenactment in this paper, i.e., driving a cartoon image with the video of a real person and vice versa. Recently, many works have focused on one-shot talking face generation to drive a portrait with a real video, i.e., within-domain reenactment. Straightforwardly applying those methods to cross-domain animation will cause inaccurate expression transfer, blur effects, and even apparent artifacts due to the domain shift between cartoon and real faces. Only a few works attempt to settle cross-domain face reenactment. The most related work AnimeCeleb requires constructing a dataset with pose vector and cartoon image pairs by animating 3D characters, which makes it inapplicable anymore if no paired data is available. In this paper, we propose a novel method for cross-domain reenactment without paired data. Specifically, we propose a transformer-based framework to align the motions from different domains into a common latent space where motion transfer is conducted via latent code addition. Two domain-specific motion encoders and two learnable motion base memories are used to capture domain properties. A source query transformer and a driving one are exploited to project domain-specific motion to the canonical space. The edited motion is projected back to the domain of the source with a transformer. Moreover, since no paired data is provided, we propose a novel cross-domain training scheme using data from two domains with the designed analogy constraint. Besides, we contribute a cartoon dataset in Disney style. Extensive evaluations demonstrate the superiority of our method over competing methods.
****
LISTER: Neighbor Decoding for Length-Insensitive Scene Text Recognition
Changxu Cheng, Peng Wang, Cheng Da, Qi Zheng, Cong Yao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19541-19551
The diversity in length constitutes a significant characteristic of text. Due to the long-tail distribution of text lengths, most existing methods for scene text recognition (STR) only work well on short or seen-length text, lacking the capability of recognizing longer text or performing length extrapolation. This is a crucial issue, since the lengths of the text to be recognized are usually not given in advance in real-world applications, but it has not been adequately investigated in previous works. Therefore, we propose in this paper a method called Length-Insensitive Scene TExt Recognizer (LISTER), which remedies the limitation regarding the robustness to various text lengths. Specifically, a Neighbor Decoder is proposed to obtain accurate character attention maps with the assistance of a novel neighbor matrix regardless of the text lengths. Besides, a Feature Enhancement Module is devised to model the long-range dependency with low computation cost, which is able to perform iterations with the neighbor decoder to enhance the feature map progressively. To the best of our knowledge, we are the first to achieve effective length-insensitive scene text recognition. Extensive experiments demonstrate that the proposed LISTER algorithm exhibits obvious superiority on long text recognition and the ability for length extrapolation, while comparing favourably with the previous state-of-the-art methods on standard benchmarks for STR (mainly short text).
****
Proxy Anchor-based Unsupervised Learning for Continuous Generalized Category Discovery
Hyungmin Kim, Sungho Suh, Daehwan Kim, Daun Jeong, Hansang Cho, Junmo Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16688-16697
Recent advances in deep learning have significantly improved the performance of various computer vision applications. However, discovering novel categories in an incremental learning scenario remains a challenging problem due to the lack of prior knowledge about the number and nature of new categories. Existing methods

for novel category discovery are limited by their reliance on labeled datasets and prior knowledge about the number of novel categories and the proportion of novel samples in the batch. To address the limitations and more accurately reflect real-world scenarios, in this paper, we propose a novel unsupervised class incremental learning approach for discovering novel categories on unlabeled sets without prior knowledge. The proposed method fine-tunes the feature extractor and proxy anchors on labeled sets, then splits samples into old and novel categories and clusters on the unlabeled dataset. Furthermore, the proxy anchors-based exemplar generates representative category vectors to mitigate catastrophic forgetting. Experimental results demonstrate that our proposed approach outperforms the state-of-the-art methods on fine-grained datasets under real-world scenarios.
****

Distribution-Aware Prompt Tuning for Vision-Language Models

Eulrang Cho, Jooyeon Kim, Hyunwoo J Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22004-22013

Pre-trained vision-language models (VLMs) have shown impressive performance on various downstream tasks by utilizing knowledge learned from large data. In general, the performance of VLMs on target tasks can be further improved by prompt tuning, which adds context to the input image or text. By leveraging data from target tasks, various prompt-tuning methods have been studied in the literature. A key to prompt tuning is the feature space alignment between two modalities via learnable vectors with model parameters fixed. We observed that the alignment becomes more effective when embeddings of each modality are 'well-arranged' in the latent space. Inspired by this observation, we proposed distribution-aware prompt tuning (DAPT) for vision-language models, which is simple yet effective. Specifically, the prompts are learned by maximizing inter-dispersion, the distance between classes, as well as minimizing the intra-dispersion measured by the distance between embeddings from the same class. Our extensive experiments on 11 benchmark datasets demonstrate that our method significantly improves generalizability. The code is available at https://github.com/mlvlab/DAPT.
****

Learning Rain Location Prior for Nighttime Deraining

Fan Zhang, Shaodi You, Yu Li, Ying Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13148-13157

Rain can significantly degrade image quality and visibility, making deraining a critical area of research in computer vision. Despite recent progress in learning-based deraining methods, there is a lack of focus on nighttime deraining due to the unique challenges posed by non-uniform local illuminations from artificial light sources. Rain streaks in these scenes have diverse appearances that are tightly related to their relative positions to light sources, making it difficult for existing deraining methods to effectively handle them. In this paper, we highlight the importance of rain streak location information in nighttime deraining. Specifically, we propose a Rain Location Prior (RLP) that is learned implicitly from rainy images using a recurrent residual model. This learned prior contains location information of rain streaks and, when injected into deraining models, can significantly improve their performance. To further improve the effectiveness of the learned prior, we also propose a Rain Prior Injection Module (RPIM) to modulate the prior before injection, increasing the importance of features within rain streak areas. Experimental results demonstrate that our approach outperforms existing state-of-the-art methods by about 1dB and effectively improves the performance of deraining models. We also evaluate our method on real night rainy images to show the capability to handle real scenes with fully synthetic data for training. Our method represents a significant step forward in the area of nighttime deraining and highlights the importance of location information in this challenging problem. The code is publicly available at https://github.com/zkawfanx/RLP.
****

FBLNet: FeedBack Loop Network for Driver Attention Prediction

Yilong Chen, Zhixiong Nan, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13371-13380

The problem of predicting driver attention from the driving perspective is gaining increasing research focus due to its remarkable significance for autonomous driving and assisted driving systems. The driving experience is extremely important for safe driving, a skilled driver is able to effortlessly predict oncoming danger (before it becomes salient) based on the driving experience and quickly pay attention to the corresponding zones. However, the nonobjective driving experience is difficult to model, so a mechanism simulating the driver experience accumulation procedure is absent in existing methods, and the current methods usually follow the technique line of saliency prediction methods to predict driver attention. In this paper, we propose a FeedBack Loop Network (FBLNet), which attempts to model the driving experience accumulation procedure. By over-and-over iterations, FBLNet generates the incremental knowledge that carries rich historically-accumulative and long-term temporal information. The incremental knowledge in our model is like the driving experience of humans. Under the guidance of the incremental knowledge, our model fuses the CNN feature and Transformer feature that are extracted from the input image to predict driver attention. Our model exhibits a solid advantage over existing methods, achieving an outstanding performance improvement on two driver attention benchmark datasets.
****
Source-free Domain Adaptive Human Pose Estimation
Qucheng Peng, Ce Zheng, Chen Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4826-4836
Human Pose Estimation (HPE) is widely used in various fields, including motion analysis, healthcare, and virtual reality. However, the great expenses of labeled real-world datasets present a significant challenge for HPE. To overcome this, one approach is to train HPE models on synthetic datasets and then perform domain adaptation (DA) on real-world data. Unfortunately, existing DA methods for HPE neglect data privacy and security by using both source and target data in the adaptation process.

To this end, we propose a new task, named source-free domain adaptive HPE, which aims to address the challenges of cross-domain learning of HPE without access to source data during the adaptation process. We further propose a novel framework that consists of three models: source model, intermediate model, and target model, which explores the task from both source-protect and target-relevant perspectives. The source-protect module preserves source information more effectively while resisting noise, and the target-relevant module reduces the sparsity of spatial representations by building a novel spatial probability space, and pose-specific contrastive learning and information maximization are proposed on the basis of this space. Comprehensive experiments on several domain adaptive HPE benchmarks show that the proposed method outperforms existing approaches by a considerable margin. The codes are available at https://github.com/davidpengucf/SFDAHPE.
****
Video Anomaly Detection via Sequentially Learning Multiple Pretext Tasks
Chenrui Shi, Che Sun, Yuwei Wu, Yunde Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10330-10340
Learning multiple pretext tasks is a popular approach to tackle the nonalignment problem in unsupervised video anomaly detection. However, the conventional learning method of simultaneously learning multiple pretext tasks, is prone to sub-optimal solutions, incurring sharp performance drops. In this paper, we propose to sequentially learn multiple pretext tasks according to their difficulties in an ascending manner to improve the performance of anomaly detection. The core idea is to relax the learning objective by starting with easy pretext tasks in the early stage and gradually refine it by involving more challenging pretext tasks later on. In this way, our method is able to reduce the difficulties of learning and avoid converging to sub-optimal solutions. Specifically, we design a tailored sequential learning order for three widely-used pretext tasks. It starts with frame prediction task, then moves on to frame reconstruction task and last ends with frame-order classification task. We further introduce a new contrastive lo

ss which makes the learned representations of normality more discriminative by pushing normal and pseudo-abnormal samples apart. Extensive experiments on three datasets demonstrate the effectiveness of our method.
****

SlaBins: Fisheye Depth Estimation using Slanted Bins on Road Environments
Jongsung Lee, Gyeongsu Cho, Jeongin Park, Kyongjun Kim, Seongoh Lee, Jung-Hee Kim, Seong-Gyun Jeong, Kyungdon Joo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8765-8774
Although 3D perception for autonomous vehicles has focused on frontal-view information, more than half of fatal accidents occur due to side impacts in practice (e.g., T-bone crash). Motivated by this fact, we investigate the problem of side-view depth estimation, especially for monocular fisheye cameras, which provide wide FoV information. However, since fisheye cameras head road areas, it observes road areas mostly and results in severe distortion on object areas, such as vehicles or pedestrians. To alleviate these issues, we propose a new fisheye depth estimation network, SlaBins, that infers an accurate and dense depth map based on a geometric property of road environments; most objects are standing (i.e., orthogonal) on the road environments. Concretely, we introduce a slanted multi-cylindrical image (MCI) representation, which allows us to describe a distance as a radius to a cylindrical layer orthogonal to the ground regardless of the camera viewing direction. Based on the slanted MCI, we estimate a set of adaptive bins and a per-pixel probability map for depth estimation. Then by combining it with the estimated slanted angle of viewing direction, we directly infer a dense and accurate depth map for fisheye cameras. Experiments demonstrate that SlaBins outperforms the state-of-the-art methods in both qualitative and quantitative evaluation on the SynWoodScape and KITTI-360 depth datasets.
****

DOT: A Distillation-Oriented Trainer
Borui Zhao, Quan Cui, Renjie Song, Jiajun Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6189-6198
Knowledge distillation transfers knowledge from a large model to a small one via task and distillation losses. In this paper, we observe a trade-off between task and distillation losses, i.e., introducing distillation loss limits the convergence of task loss. We believe that the trade-off results from the insufficient optimization of distillation loss. The reason is: The teacher has a lower task loss than the student, and a lower distillation loss drives the student more similar to the teacher, then a better-converged task loss could be obtained. To break the trade-off, we propose the Distillation-Oriented Trainer (DOT). DOT separately considers gradients of task and distillation losses, then applies a larger momentum to distillation loss to accelerate its optimization. We empirically prove that DOT breaks the trade-off, i.e., both losses are sufficiently optimized. Extensive experiments validate the superiority of DOT. Notably, DOT achieves a +2.59% accuracy improvement on ImageNet-1k for the ResNet50-MobileNetV1 pair. Conclusively, DOT greatly benefits the student's optimization properties in terms of loss convergence and model generalization. Code will be made publicly available.
****

Neural Collage Transfer: Artistic Reconstruction via Material Manipulation
Ganghun Lee, Minji Kim, Yunsu Lee, Minsu Lee, Byoung-Tak Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2394-2405
Collage is a creative art form that uses diverse material scraps as a base unit to compose a single image.
Although pixel-wise generation techniques can reproduce a target image in collage style, it is not a suitable method due to the solid stroke-by-stroke nature of the collage form.
While some previous works for stroke-based rendering produced decent sketches and paintings, collages have received much less attention in research despite their popularity as a style.
In this paper, we propose a method for learning to make collages via reinforcem

ent learning without the need for demonstrations or collage artwork data.
We design the collage Markov Decision Process (MDP), which allows the agent to handle various materials and propose a model-based soft actor-critic to mitigate the agent's training burden derived from the sophisticated dynamics of collage. Moreover, we devise additional techniques such as active material selection and complexity-based multi-scale collage to handle target images at any size and enhance the results' aesthetics by placing relatively more scraps in areas of high complexity.
Experimental results show that the trained agent appropriately selected and pasted materials to regenerate the target image into a collage and obtained a higher evaluation score on content and style than pixel-wise generation methods. Code is available at https://github.com/northadventure/CollageRL.
****
Fantasia3D: Disentangling Geometry and Appearance for High-quality Text-to-3D Content Creation
Rui Chen, Yongwei Chen, Ningxin Jiao, Kui Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22246-22256
Automatic 3D content creation has achieved rapid progress recently due to the availability of pre-trained, large language models and image diffusion models, forming the emerging topic of text-to-3D content creation. Existing text-to-3D methods commonly use implicit scene representations, which couple the geometry and appearance via volume rendering and are suboptimal in terms of recovering finer geometries and achieving photorealistic rendering; consequently, they are less effective for generating high-quality 3D assets. In this work, we propose a new method of Fantasia3D for high-quality text-to-3D content creation. Key to Fantasia3D is the disentangled modeling and learning of geometry and appearance. For geometry learning, we rely on a hybrid scene representation, and propose to encode surface normal extracted from the representation as the input of the image diffusion model. For appearance modeling, we introduce the spatially varying bidirectional reflectance distribution function (BRDF) into the text-to-3D task, and learn the surface material for photorealistic rendering of the generated surface. Our disentangled framework is more compatible with popular graphics engines, supporting relighting, editing, and physical simulation of the generated 3D assets. We conduct thorough experiments that show the advantages of our method over existing ones under different text-to-3D task settings. Project page and source codes: https://fantasia3d.github.io/.
****
MagicFusion: Boosting Text-to-Image Generation Performance by Fusing Diffusion Models
Jing Zhao, Heliang Zheng, Chaoyue Wang, Long Lan, Wenjing Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22592-22602
The advent of open-source AI communities has produced a cornucopia of powerful text-guided diffusion models that are trained on various datasets. While few explorations have been conducted on ensembling such models to combine their strengths. In this work, we propose a simple yet effective method called Saliency-aware Noise Blending (SNB) that can empower the fused text-guided diffusion models to achieve more controllable generation. Specifically, we experimentally find that the responses of classifier-free guidance are highly related to the saliency of generated images. Thus we propose to trust different models in their areas of expertise by blending the predicted noises of two diffusion models in a saliency-aware manner. SNB is training-free and can be completed within a DDIM sampling process. Additionally, it can automatically align the semantics of two noise spaces without requiring additional annotations such as masks. Extensive experiments show the impressive effectiveness of SNB in various applications. The project page is available at https://magicfusion.github.io.
****
UCF: Uncovering Common Features for Generalizable Deepfake Detection
Zhiyuan Yan, Yong Zhang, Yanbo Fan, Baoyuan Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22412-22423

Deepfake detection remains a challenging task due to the difficulty of generalizing to new types of forgeries. This problem primarily stems from the overfitting of existing detection methods to forgery-irrelevant features and method-specific patterns. The latter has been rarely studied and not well addressed by previous works. This paper presents a novel approach to address the two types of overfitting issues by uncovering common forgery features. Specifically, we first propose a disentanglement framework that decomposes image information into three distinct components: forgery-irrelevant, method-specific forgery, and common forgery features. To ensure the decoupling of method-specific and common forgery features, a multi-task learning strategy is employed, including a multi-class classification that predicts the category of the forgery method and a binary classification that distinguishes the real from the fake. Additionally, a conditional decoder is designed to utilize forgery features as a condition along with forgery-irrelevant features to generate reconstructed images. Furthermore, a contrastive regularization technique is proposed to encourage the disentanglement of the common and specific forgery features. Ultimately, we only utilize the common forgery features for the purpose of generalizable deepfake detection. Extensive evaluations demonstrate that our framework can perform superior generalization than current state-of-the-art methods.
****
March in Chat: Interactive Prompting for Remote Embodied Referring Expression
Yanyuan Qiao, Yuankai Qi, Zheng Yu, Jing Liu, Qi Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15758-15767
Many Vision-and-Language Navigation (VLN) tasks have been proposed in recent years, from room-based to object-based and indoor to outdoor. The REVERIE (Remote Embodied Referring Expression) is interesting since it only provides high-level instructions to the agent, which are closer to human commands in practice. Nevertheless, this poses more challenges than other VLN tasks since it requires agents to infer a navigation plan only based on a short instruction. Large Language Models (LLMs) show great potential in robot action planning by providing proper prompts. Still, this strategy has not been explored under the REVERIE settings. There are several new challenges. For example, the LLM should be environment-aware so that the navigation plan can be adjusted based on the current visual observation. Moreover, the LLM planned actions should be adaptable to the much larger and more complex REVERIE environment. This paper proposes a March-in-Chat (MiC) model that can talk to the LLM on the fly and plan dynamically based on a newly proposed Room-and-Object Aware Scene Perceiver (ROASP). Our MiC model outperforms the previous state-of-the-art by large margins by SPL and RGSPL metrics on the REVERIE benchmark. The source code is available at https://github.com/YanyuanQiao/MiC
****
Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation
Fabian Deuser, Konrad Habel, Norbert Oswald; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16847-16856
Cross-View Geo-Localisation is still a challenging task where additional modules, specific pre-processing or zooming strategies are necessary to determine accurate positions of images. Since different views have different geometries, pre-processing like polar transformation helps to merge them. However, this results in distorted images which then have to be rectified. Adding hard negatives to the training batch could improve the overall performance but with the default loss functions in geo-localisation it is difficult to include them.
In this article, we present a simplified but effective architecture based on contrastive learning with symmetric InfoNCE loss that outperforms current state-of-the-art results. Our framework consists of a narrow training pipeline that eliminates the need of using aggregation modules, avoids further pre-processing steps and even increases the generalisation capability of the model to unknown regions. We introduce two types of sampling strategies for hard negatives. The first explicitly exploits geographically neighboring locations to provide a good starting point. The second leverages the visual similarity between the image embeddings in order to mine hard negative samples. Our work shows excellent performance

on common cross-view datasets like CVUSA, CVACT, University-1652 and VIGOR. A co
mparison between cross-area and same-area settings demonstrate the good generali
sation capability of our model.
****
Novel Scenes & Classes: Towards Adaptive Open-set Object Detection
Wuyang Li, Xiaoqing Guo, Yixuan Yuan; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2023, pp. 15780-15790
Domain Adaptive Object Detection (DAOD) transfers an object detector to a novel
domain free of labels. However, in the real world, besides encountering novel sc
enes, novel domains always contain novel-class objects de facto, which are ignor
ed in existing research. Thus, we formulate and study a more practical setting,
Adaptive Open-set Object Detection (AOOD), considering both novel scenes and cla
sses. Directly combing off-the-shelled cross-domain and open-set approaches is s
ub-optimal since their low-order dependence, e.g., the confidence score, is insu
fficient for the AOOD with two dimensions of novel information. To address this,
 we propose a novel Structured mOtif MAtching (SOMA) framework for AOOD, which m
odels the high-order relation with motifs, i.e., a statistically significant sub
graph, and formulates AOOD solution as motif matching to learn with high-order p
atterns. In a nutshell, SOMA consists of Structure-aware Novel-class Learning (S
NL) and Structure-aware Transfer Learning (STL). As for SNL, we establish an ins
tance-oriented graph to capture the class-independent object feature hidden in d
ifferent base classes. Then, a high-order metric is proposed to match the most s
ignificant motif as high-order patterns, serving for motif-guided novel-class le
arning. In STL, we set up a semantic-oriented graph to model the class-dependent
 relation across domains, and match unlabelled objects with high-order motifs to
 align the cross-domain distribution with structural awareness. Extensive experi
ments demonstrate that the proposed SOMA achieves state-of-the-art performance.
Codes will be released publicly for further study.
****
LIMITR: Leveraging Local Information for Medical Image-Text Representation
Gefen Dawidowicz, Elad Hirsch, Ayellet Tal; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 21165-21173
Medical imaging analysis plays a critical role in the diagnosis and treatment of
 various medical conditions. This paper focuses on chest X-ray images and their
corresponding radiological reports. It presents a new model that learns a joint
X-ray image & report representation. The model is based on a novel alignment sch
eme between the visual data and the text, which takes into account both local an
d global information. Furthermore, the model integrates domain-specific informat
ion of two types -- lateral images and the consistent visual structure of chest
images. Our representation is shown to benefit three types of retrieval tasks: t
ext-image retrieval, class-based retrieval, and phrase-grounding.
****
Multi-task View Synthesis with Neural Radiance Fields
Shuhong Zheng, Zhipeng Bao, Martial Hebert, Yu-Xiong Wang; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21538-21549
Multi-task visual learning is a critical aspect of computer vision. Current rese
arch, however, predominantly concentrates on the multi-task dense prediction set
ting, which overlooks the intrinsic 3D world and its multi-view consistent struc
tures, and lacks the capacity for versatile imagination. In response to these li
mitations, we present a novel problem setting -- multi-task view synthesis (MTVS
), which reinterprets multi-task prediction as a set of novel-view synthesis tas
ks for multiple scene properties, including RGB. To tackle the MTVS problem, we
propose MuvieNeRF, a framework that incorporates both multi-task and cross-view
knowledge to simultaneously synthesize multiple scene properties. MuvieNeRF inte
grates two key modules, the Cross-Task Attention (CTA) and Cross-View Attention
(CVA) modules, enabling the efficient use of information across multiple views a
nd tasks. Extensive evaluations on both synthetic and realistic benchmarks demon
strate that MuvieNeRF is capable of simultaneously synthesizing different scene
properties with promising visual quality, even outperforming conventional discri
minative models in various settings. Notably, we show that MuvieNeRF exhibits un

iversal applicability across a range of NeRF backbones. Our code is available at https://github.com/zsh2000/MuvieNeRF.
****
Informative Data Mining for One-Shot Cross-Domain Semantic Segmentation
Yuxi Wang, Jian Liang, Jun Xiao, Shuqi Mei, Yuran Yang, Zhaoxiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1064-1074
Contemporary domain adaptation offers a practical solution for achieving cross-domain transfer of semantic segmentation between labelled source data and unlabeled target data. These solutions have gained significant popularity; however, they require the model to be retrained when the test environment changes. This can result in unbearable costs in certain applications due to the time-consuming training process and concerns regarding data privacy. One-shot domain adaptation methods attempt to overcome these challenges by transferring the pre-trained source model to the target domain using only one target data. Despite this, the referring style transfer module still faces issues with computation cost and over-fitting problems.

To address this problem, we propose a novel framework called Informative Data Mining (IDM) that enables efficient one-shot domain adaptation for semantic segmentation. Specifically, IDM provides an uncertainty-based selection criterion to identify the most informative samples, which facilitates quick adaptation and reduces redundant training. We then perform a model adaptation method using these selected samples, which includes patch-wise mixing and prototype-based information maximization to update the model. This approach effectively enhances adaptation and mitigates the overfitting problem.

In general, we provide empirical evidence of the effectiveness and efficiency of IDM. Our approach outperforms existing methods and achieves a new state-of-the-art one-shot performance of 56.7%/55.4% on the GTA5/SYNTHIA to Cityscapes adaptation tasks, respectively. The code will be released at https://github.com/yxiwang/IDM.
****
Efficient Unified Demosaicing for Bayer and Non-Bayer Patterned Image Sensors
Haechang Lee, Dongwon Park, Wongi Jeong, Kijeong Kim, Hyunwoo Je, Dongil Ryu, Se Young Chun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12750-12759
As the physical size of recent CMOS image sensors (CIS) gets smaller, the latest mobile cameras are adopting unique non-Bayer color filter array (CFA) patterns (e.g., Quad, Nona, QxQ), which consist of homogeneous color units with adjacent pixels. These non-Bayer sensors are superior to conventional Bayer CFA thanks to their changeable pixel-bin sizes for different light conditions, but may introduce visual artifacts during demosaicing due to their inherent pixel pattern structures and sensor hardware characteristics. Previous demosaicing methods have primarily focused on fixed pixel-bin sizes of Bayer CFA, requiring specialized reconstruction methods for non-Bayer patterned CIS and executing multiple CFA modes depending on lighting conditions. In this work, we propose an efficient unified demosaicing method that can be applied to both conventional Bayer RAW and various non-Bayer CFAs' RAW data in different operation modes. Our Knowledge Learning-based demosaicing model for Adaptive Patterns, namely KLAP, utilizes CFA-adaptive filters for only 1% key filters in the network for each CFA, but still manages to effectively demosaic all the CFAs, yielding comparable performance to the large-scale models. Furthermore, by employing meta-learning during inference (KLAP-M), our model is able to eliminate unknown sensor-generic artifacts in real RAW data, effectively bridging the gap between synthetic images and real sensor RAW. Our KLAP and KLAP-M methods achieved state-of-the-art demosaicing performance in both synthetic and real RAW data of Bayer and non-Bayer CFAs.
****
Visual Traffic Knowledge Graph Generation from Scene Images
Yunfei Guo, Fei Yin, Xiao-hui Li, Xudong Yan, Tao Xue, Shuqi Mei, Cheng-Lin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21604-21613

Although previous works on traffic scene understanding have achieved great succe ss, most of them stop at a lowlevel perception stage, such as road segmentation and lane detection, and few concern high-level understanding. In this paper, we present Visual Traffic Knowledge Graph Generation (VTKGG), a new task for in-dep th traffic scene understanding that tries to extract multiple kinds of informati on and integrate them into a knowledge graph. To achieve this goal, we first int roduce a large dataset named CASIATencent Road Scene dataset (RS10K) with compre hensive annotations to support related research. Secondly, we propose a novel tr affic scene parsing architecture containing a Hierarchical Graph ATtention netwo rk (HGAT) to analyze the heterogeneous elements and their complicated relations in traffic scene images. By hierarchizing the heterogeneous graph and equipping it with cross-level links, our approach exploits the correlation among various e lements completely and acquires accurate relations. The experimental results sho w that our method can effectively generate visual traffic knowledge graphs and a chieve state-of-the-art performance.
****
Householder Projector for Unsupervised Latent Semantics Discovery
Yue Song, Jichao Zhang, Nicu Sebe, Wei Wang; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2023, pp. 7712-7722
Generative Adversarial Networks (GANs), especially the recent style-based genera tors (StyleGANs), have versatile semantics in the structured latent space. Laten t semantics discovery methods emerge to move around the latent code such that on ly one factor varies during the traversal. Recently, an unsupervised method prop osed a promising direction to directly use the eigenvectors of the projection ma trix that maps latent codes to features as the interpretable directions. However , one overlooked fact is that the projection matrix is non-orthogonal and the nu mber of eigenvectors is too large. The non-orthogonality would entangle semantic attributes in the top few eigenvectors, and the large dimensionality might resu lt in meaningless variations among the directions even if the matrix is orthogon al. To avoid these issues, we propose Householder Projector, a flexible and gene ral low-rank orthogonal matrix representation based on Householder transformatio ns, to parameterize the projection matrix. The orthogonality guarantees that the eigenvectors correspond to disentangled interpretable semantics, while the low- rank property encourages that each identified direction has meaningful variation s. We integrate our projector into pre-trained StyleGAN2/StyleGAN3 and evaluate the models on several benchmarks. Within only 1% of the original training steps for fine-tuning, our projector helps StyleGANs to discover more disentangled and precise semantic attributes without sacrificing image fidelity.
****
Spatially-Adaptive Feature Modulation for Efficient Image Super-Resolution
Long Sun, Jiangxin Dong, Jinhui Tang, Jinshan Pan; Proceedings of the IEEE/CVF I nternational Conference on Computer Vision (ICCV), 2023, pp. 13190-13199
Although deep learning-based solutions have achieved impressive reconstruction p erformance in image super-resolution (SR), these models are generally large, wit h complex architectures, making them incompatible with low-power devices with ma ny computational and memory constraints. To overcome these challenges, we propos e a spatially-adaptive feature modulation (SAFM) mechanism for efficient SR desi gn. In detail, the SAFM layer uses independent computations to learn multi-scale feature representations and aggregates these features for dynamic spatial modul ation. As the SAFM prioritizes exploiting non-local feature dependencies, we fur ther introduce a convolutional channel mixer (CCM) to encode local contextual in formation and mix channels simultaneously. Extensive experimental results show t hat the proposed method is 3x smaller than state-of-the-art efficient SR methods , e.g., IMDN, and yields comparable performance with much less memory usage. Our source codes and pre-trained models are available at: https://github.com/sunny2 109/SAFMN.
****
Unsupervised Image Denoising in Real-World Scenarios via Self-Collaboration Para llel Generative Adversarial Branches
Xin Lin, Chao Ren, Xiao Liu, Jie Huang, Yinjie Lei; Proceedings of the IEEE/CVF

International Conference on Computer Vision (ICCV), 2023, pp. 12642-12652

Deep learning methods have shown remarkable performance in image denoising, particularly when trained on large-scale paired datasets. However, acquiring such paired datasets for real-world scenarios poses a significant challenge. Although unsupervised approaches based on generative adversarial networks (GANs) offer a promising solution for denoising without paired datasets, they are difficult in surpassing the performance limitations of conventional GAN-based unsupervised frameworks without significantly modifying existing structures or increasing the computational complexity of denoisers. To address this problem, we propose a self-collaboration (SC) strategy for multiple denoisers. This strategy can achieve significant performance improvement without increasing the inference complexity of the GAN-based denoising framework. Its basic idea is to iteratively replace the previous less powerful denoiser in the filter-guided noise extraction module with the current powerful denoiser. This process generates better synthetic clean-noisy image pairs, leading to a more powerful denoiser for the next iteration. In addition, we propose a baseline method that includes parallel generative adversarial branches with complementary "self-synthesis" and "unpaired-synthesis" constraints. This baseline ensures the stability and effectiveness of the training network. The experimental results demonstrate the superiority of our method over state-of-the-art unsupervised methods.

****

Bayesian Optimization Meets Self-Distillation

HyunJae Lee, Heon Song, Hyeonsoo Lee, Gi-hyeon Lee, Suyeong Park, Donggeun Yoo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1696-1705

Bayesian optimization (BO) has contributed greatly to improving model performance by suggesting promising hyperparameter configurations iteratively based on observations from multiple training trials. However, only partial knowledge (i.e., the measured performances of trained models and their hyperparameter configurations) from previous trials is transferred. On the other hand, Self-Distillation (SD) only transfers partial knowledge learned by the task model itself. To fully leverage the various knowledge gained from all training trials, we propose the BOSS framework, which combines BO and SD. BOSS suggests promising hyperparameter configurations through BO and carefully selects pre-trained models from previous trials for SD, which are otherwise abandoned in the conventional BO process. BOSS achieves significantly better performance than both BO and SD in a wide range of tasks including general image classification, learning with noisy labels, semi-supervised learning, and medical image analysis tasks. Our code is available at https://github.com/sooperset/boss.

****

No Fear of Classifier Biases: Neural Collapse Inspired Federated Learning with Synthetic and Fixed Classifier

Zexi Li, Xinyi Shang, Rui He, Tao Lin, Chao Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5319-5329

Data heterogeneity is an inherent challenge that hinders the performance of federated learning (FL). Recent studies have identified the biased classifiers of local models as the key bottleneck. Previous attempts have used classifier calibration after FL training, but this approach falls short in improving the poor feature representations caused by training-time classifier biases. Resolving the classifier bias dilemma in FL requires a full understanding of the mechanisms behind the classifier. Recent advances in neural collapse have shown that the classifiers and feature prototypes under perfect training scenarios collapse into an optimal structure called simplex equiangular tight frame (ETF). Building on this neural collapse insight, we propose a solution to the FL's classifier bias problem by utilizing a synthetic and fixed ETF classifier during training. The optimal classifier structure enables all clients to learn unified and optimal feature representations even under extremely heterogeneous data. We devise several effective modules to better adapt the ETF structure in FL, achieving both high generalization and personalization. Extensive experiments demonstrate that our method achieves state-of-the-art performances on CIFAR-10, CIFAR-100, and Tiny-ImageNet.

The code is available at https://github.com/ZexiLee/ICCV-2023-FedETF.
****
MemorySeg: Online LiDAR Semantic Segmentation with a Latent Memory
Enxu Li, Sergio Casas, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 745-754
Semantic segmentation of LiDAR point clouds has been widely studied in recent years, with most existing methods focusing on tackling this task using a single scan of the environment. However, leveraging the temporal stream of observations can provide very rich contextual information on regions of the scene with poor visibility (e.g., occlusions) or sparse observations (e.g., at long range), and can help reduce redundant computation frame after frame. In this paper, we tackle the challenge of exploiting the information from the past frames to improve the predictions of the current frame in an online fashion. To address this challenge, we propose a novel framework for semantic segmentation of a temporal sequence of LiDAR point clouds that utilizes a memory network to store, update and retrieve past information. Our framework also includes a novel regularizer that penalizes prediction variations in the neighborhood of the point cloud. Prior works have attempted to incorporate memory in range view representations for semantic segmentation, but these methods fail to handle occlusions and the range view representation of the scene changes drastically as agents nearby move. Our proposed framework overcomes these limitations by building a sparse 3D latent representation of the surroundings. We evaluate our method on SemanticKITTI, nuScenes, and PandaSet. Our experiments demonstrate the effectiveness of the proposed framework compared to the state-of-the-art. For more information, visit the project website: https://waabi.ai/research/memoryseg.
****
Hashing Neural Video Decomposition with Multiplicative Residuals in Space-Time
Cheng-Hung Chan, Cheng-Yang Yuan, Cheng Sun, Hwann-Tzong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7743-7753
We present a video decomposition method that facilitates layer-based editing of videos with spatiotemporally varying lighting and motion effects.
Our neural model decomposes an input video into multiple layered representations, each comprising a 2D texture map, a mask for the original video, and a multiplicative residual characterizing the spatiotemporal variations in lighting conditions.
A single edit on the texture maps can be propagated to the corresponding locations in the entire video frames while preserving other contents' consistencies.
Our method efficiently learns the layer-based neural representations of a 1080p video in 25s per frame via coordinate hashing and allows real-time rendering of the edited result at 71 fps on a single GPU.
Qualitatively, we run our method on various videos to show its effectiveness in generating high-quality editing effects. Quantitatively, we propose to adopt feature-tracking evaluation metrics for objectively assessing the consistency of video editing.
Project page: https://lightbulb12294.github.io/hashing-nvd/
****
Multimodal Variational Auto-encoder based Audio-Visual Segmentation
Yuxin Mao, Jing Zhang, Mochu Xiang, Yiran Zhong, Yuchao Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 954-965
We propose an Explicit Conditional Multimodal Variational Auto-Encoder (ECMVAE) for audio-visual segmentation (AVS), aiming to segment sound sources in the video sequence. Existing AVS methods focus on implicit feature fusion strategies, where models are trained to fit the discrete samples in the dataset. With a limited and less diverse dataset, the resulting performance is usually unsatisfactory. In contrast, we address this problem from an effective representation learning perspective, aiming to model the contribution of each modality explicitly. Specifically, we find that audio contains critical category information of the sound producers, and visual data provides candidate sound producer(s). Their shared information corresponds to the target sound producer(s) shown in the visual data.

In this case, cross-modal shared representation learning is especially important for AVS. To achieve this, our ECMVAE factorizes the representations of each modality with a modality-shared representation and a modality-specific representation. An orthogonality constraint is applied between the shared and specific representations to maintain the exclusive attribute of the factorized latent code. Further, a mutual information maximization regularizer is introduced to achieve extensive exploration of each modality. Quantitative and qualitative evaluations on the AVSBench demonstrate the effectiveness of our approach, leading to a new state-of-the-art for AVS, with a 3.84 mIOU performance leap on the challenging MS3 subset for multiple sound source segmentation. Code and pre-train model will release to provide full details of our proposed method.
****
DynaMITe: Dynamic Query Bootstrapping for Multi-object Interactive Segmentation Transformer
Amit Kumar Rana, Sabarinath Mahadevan, Alexander Hermans, Bastian Leibe; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1043-1052
Most state-of-the-art instance segmentation methods rely on large amounts of pixel-precise ground-truth annotations for training, which are expensive to create. Interactive segmentation networks help generate such annotations based on an image and the corresponding user interactions such as clicks. Existing methods for this task can only process a single instance at a time and each user interaction requires a full forward pass through the entire deep network. We introduce a more efficient approach, called DynaMITe, in which we represent user interactions as spatio-temporal queries to a Transformer decoder with a potential to segment multiple object instances in a single iteration. Our architecture also alleviates any need to re-compute image features during refinement, and requires fewer interactions for segmenting multiple instances in a single image when compared to other methods. DynaMITe achieves state-of-the-art results on multiple existing interactive segmentation benchmarks, and also on the new multi-instance benchmark that we propose in this paper.
****
FRAug: Tackling Federated Learning with Non-IID Features via Representation Augmentation
Haokun Chen, Ahmed Frikha, Denis Krompass, Jindong Gu, Volker Tresp; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4849-4859
Federated Learning (FL) is a decentralized machine learning paradigm, in which multiple clients collaboratively train neural networks without centralizing their local data, and hence preserve data privacy. However, real-world FL applications usually encounter challenges arising from distribution shifts across the local datasets of individual clients. These shifts may drift the global model aggregation or result in convergence to deflected local optimum. While existing efforts have addressed distribution shifts in the label space, an equally important challenge remains relatively unexplored. This challenge involves situations where the local data of different clients indicate identical label distributions but exhibit divergent feature distributions. This issue can significantly impact the global model performance in the FL framework. In this work, we propose Federated Representation Augmentation (FRAug) to resolve this practical and challenging problem. FRAug optimizes a shared embedding generator to capture client consensus. Its output synthetic embeddings are transformed into client-specific by a locally optimized RTNet to augment the training space of each client. Our empirical evaluation on three public benchmarks and a real-world medical dataset demonstrates the effectiveness of the proposed method, which substantially outperforms the current state-of-the-art FL methods for feature distribution shifts, including PartialFed and FedBN.
****
Homography Guided Temporal Fusion for Road Line and Marking Segmentation
Shan Wang, Chuong Nguyen, Jiawei Liu, Kaihao Zhang, Wenhan Luo, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, Hongdong Li; Proceedings of the IEEE/CVF Inter

national Conference on Computer Vision (ICCV), 2023, pp. 1075-1085
Reliable segmentation of road lines and markings is critical to autonomous driving. Our work is motivated by the observations that road lines and markings are (1) frequently occluded in the presence of moving vehicles, shadow, and glare and (2) highly structured with low intra-class shape variance and overall high appearance consistency. To solve these issues, we propose a Homography Guided Fusion (HomoFusion) module to exploit temporally-adjacent video frames for complementary cues facilitating the correct classification of the partially occluded road lines or markings. To reduce computational complexity, a novel surface normal estimator is proposed to establish spatial correspondences between the sampled frames, allowing the HomoFusion module to perform a pixel-to-pixel attention mechanism in updating the representation of the occluded road lines or markings. Experiments on ApolloScape, a large-scale lane mark segmentation dataset, and ApolloScape Night with artificial simulated night-time road conditions, demonstrate that our method outperforms other existing SOTA lane mark segmentation models with less than 9% of their parameters and computational complexity. We show that exploiting available camera intrinsic data and ground plane assumption for cross-frame correspondence can lead to a light-weight network with significantly improved performances in speed and accuracy. We also prove the versatility of our HomoFusion approach by applying it to the problem of water puddle segmentation and achieving SOTA performance.
****
NeuRBF: A Neural Fields Representation with Adaptive Radial Basis Functions
Zhang Chen, Zhong Li, Liangchen Song, Lele Chen, Jingyi Yu, Junsong Yuan, Yi Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4182-4194
We present a novel type of neural fields that uses general radial bases for signal representation. State-of-the-art neural fields typically rely on grid-based representations for storing local neural features and N-dimensional linear kernels for interpolating features at continuous query points. The spatial positions of their neural features are fixed on grid nodes and cannot well adapt to target signals. Our method instead builds upon general radial bases with flexible kernel position and shape, which have higher spatial adaptivity and can more closely fit target signals. To further improve the channel-wise capacity of radial basis functions, we propose to compose them with multi-frequency sinusoid functions. This technique extends a radial basis to multiple Fourier radial bases of different frequency bands without requiring extra parameters, facilitating the representation of details. Moreover, by marrying adaptive radial bases with grid-based ones, our hybrid combination inherits both adaptivity and interpolation smoothness. We carefully designed weighting schemes to let radial bases adapt to different types of signals effectively. Our experiments on 2D image and 3D signed distance field representation demonstrate the higher accuracy and compactness of our method than prior arts. When applied to neural radiance field reconstruction, our method achieves state-of-the-art rendering quality, with small model size and comparable training speed.
****
OmnimatteRF: Robust Omnimatte with 3D Background Modeling
Geng Lin, Chen Gao, Jia-Bin Huang, Changil Kim, Yipeng Wang, Matthias Zwicker, Ayush Saraf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23471-23480
Video matting has broad applications, from adding interesting effects to casually captured movies to assisting video production professionals.
Matting with associated effects such as shadows and reflections has also attracted increasing research activity, and methods like Omnimatte have been proposed to separate dynamic foreground objects of interest into their own layers.
However, prior works represent video backgrounds as 2D image layers, limiting their capacity to express more complicated scenes, thus hindering application to real-world videos.
In this paper, we propose a novel video matting method, OmnimatteRF, that combines dynamic 2D foreground layers and a 3D background model.

The 2D layers preserve the details of the subjects, while the 3D background robustly reconstructs scenes in real-world videos.
Extensive experiments demonstrate that our method reconstructs scenes with better quality on various videos.
****
Self-supervised Image Denoising with Downsampled Invariance Loss and Conditional Blind-Spot Network
Yeong Il Jang, Keuntek Lee, Gu Yong Park, Seyun Kim, Nam Ik Cho; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12196-12205
There have been many image denoisers using deep neural networks, which outperform conventional model-based methods by large margins. Recently, self-supervised methods have attracted attention because constructing a large real noise dataset for supervised training is an enormous burden. The most representative self-supervised denoisers are based on blind-spot networks, which exclude the receptive field's center pixel. However, excluding any input pixel is abandoning some information, especially when the input pixel at the corresponding output position is excluded. In addition, a standard blind-spot network fails to reduce real camera noise due to the pixel-wise correlation of noise, though it successfully removes independently distributed synthetic noise. Hence, to realize a more practical denoiser, we propose a novel self-supervised training framework that can remove real noise. For this, we derive the theoretic upper bound of a supervised loss where the network is guided by the downsampled blinded output. Also, we design a conditional blind-spot network (C-BSN), which selectively controls the blindness of the network to use the center pixel information. Furthermore, we exploit a random subsampler to decorrelate noise spatially, making the C-BSN free of visual artifacts that were often seen in downsample-based methods. Extensive experiments show that the proposed C-BSN achieves state-of-the-art performance on real-world datasets as a self-supervised denoiser and shows qualitatively pleasing results without any post-processing or refinement.
****
Multi-granularity Interaction Simulation for Unsupervised Interactive Segmentation
Kehan Li, Yian Zhao, Zhennan Wang, Zesen Cheng, Peng Jin, Xiangyang Ji, Li Yuan, Chang Liu, Jie Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 666-676
Interactive segmentation enables users to segment as needed by providing cues of objects, which introduces human-computer interaction for many fields, such as image editing and medical image analysis. Typically, massive and expansive pixel-level annotations are spent to train deep models by object-oriented interactions with manually labeled object masks. In this work, we reveal that informative interactions can be made by simulation with semantic-consistent yet diverse region exploration in an unsupervised paradigm. Concretely, we introduce a Multi-granularity Interaction Simulation (MIS) approach to open up a promising direction for unsupervised interactive segmentation. Drawing on the high-quality dense features produced by recent self-supervised models, we propose to gradually merge patches or regions with similar features to form more extensive regions and thus, every merged region serves as a semantic-meaningful multi-granularity proposal. By randomly sampling these proposals and simulating possible interactions based on them, we provide meaningful interaction at multiple granularities to teach the model to understand interactions. Our MIS significantly outperforms non-deep learning unsupervised methods and is even comparable with some previous deep-supervised methods without any annotation.
****
RecursiveDet: End-to-End Region-Based Recursive Object Detection
Jing Zhao, Li Sun, Qingli Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6307-6316
End-to-end region-based object detectors like Sparse R-CNN usually have multiple cascade bounding box decoding stages, which refine the current predictions according to their previous results. Model parameters within each stage are independ

ent, evolving a huge cost. In this paper, we find the general setting of decoding stages is actually redundant. By simply sharing parameters and making a recursive decoder, the detector already obtains a significant improvement. The recursive decoder can be further enhanced by positional encoding (PE) of the proposal box, which makes it aware of the exact locations and sizes of input bounding boxes, thus becoming adaptive to proposals from different stages during the recursion. Moreover, we also design centerness-based PE to distinguish the RoI feature element and dynamic convolution kernels at different positions within the bounding box. To validate the effectiveness of the proposed method, we conduct intensive ablations and build the full model on three recent mainstream region-based detectors. The RecusiveDet is able to achieve obvious performance boosts with even fewer model parameters and slightly increased computation cost.
****
Bold but Cautious: Unlocking the Potential of Personalized Federated Learning through Cautiously Aggressive Collaboration
Xinghao Wu, Xuefeng Liu, Jianwei Niu, Guogang Zhu, Shaojie Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19375-19384
Personalized federated learning (PFL) reduces the impact of non-independent and identically distributed (non-IID) data among clients by allowing each client to train a personalized model when collaborating with others. A key question in PFL is to decide which parameters of a client should be localized or shared with others. In current mainstream approaches, all layers that are sensitive to non-IID data (such as classifier layers) are generally personalized. The reasoning behind this approach is understandable, as localizing parameters that are easily influenced by non-IID data can prevent potential negative effects of collaboration. However, we believe that this approach is too conservative for collaboration. For example, for a certain client, even if its parameters are easily influenced by non-IID data, it can still benefit by sharing these parameters with clients having similar data distribution. This observation emphasizes the importance of considering not only the sensitivity to non-IID data but also the similarity of data distribution when determining which parameters should be localized in PFL. This paper introduces a novel guideline for client collaboration in PFL. Unlike existing approaches that prohibit all collaboration of sensitive parameters, our guideline allows clients to share more parameters with others, leading to improved model performance. Additionally, we propose a new PFL method named FedCAC, which employs a quantitative metric to evaluate each parameter's sensitivity to non-IID data and carefully selects collaborators based on this evaluation. Experimental results demonstrate that FedCAC enables clients to share more parameters with others, resulting in superior performance compared to state-of-the-art methods, particularly in scenarios where clients have diverse distributions.
****
ESSAformer: Efficient Transformer for Hyperspectral Image Super-resolution
Mingjin Zhang, Chi Zhang, Qiming Zhang, Jie Guo, Xinbo Gao, Jing Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23073-23084
Single hyperspectral image super-resolution (single-HSI-SR) aims to restore a high-resolution hyperspectral image from a low-resolution observation. However, the prevailing CNN-based approaches have shown limitations in building long-range dependencies and capturing interaction information between spectral features. This results in inadequate utilization of spectral information and artifacts after upsampling. To address this issue, we propose ESSAformer, an ESSA attention-embedded Transformer network for single-HSI-SR with an iterative refining structure. Specifically, we first introduce a robust and spectral-friendly similarity metric, i.e., the spectral correlation coefficient of the spectrum (SCC), to replace the original attention matrix and incorporates inductive biases into the model to facilitate training. Built upon it, we further utilize the kernelizable attention technique with theoretical support to form a novel efficient SCC-kernel-based self-attention (ESSA) and reduce attention computation to linear complexity. ESSA enlarges the receptive field for features after upsampling without bringin

g much computation and allows the model to effectively utilize spatial-spectral information from different scales, resulting in the generation of more natural high-resolution images. Without the need for pretraining on large-scale datasets, our experiments demonstrate ESSA's effectiveness in both visual quality and quantitative results. The code will be released.
****
Generative Action Description Prompts for Skeleton-based Action Recognition
Wangmeng Xiang, Chao Li, Yuxuan Zhou, Biao Wang, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10276-10285
Skeleton-based action recognition has recently received considerable attention. Current approaches to skeleton-based action recognition are typically formulated as one-hot classification tasks and do not fully exploit the semantic relations between actions. For example, "make victory sign" and "thumb up" are two actions of hand gestures, whose major difference lies in the movement of hands. This information is agnostic from the categorical one-hot encoding of action classes but could be unveiled from the action description. Therefore, utilizing action description in training could potentially benefit representation learning. In this work, we propose a Generative Action-description Prompts (GAP) approach for skeleton-based action recognition. More specifically, we employ a pre-trained large-scale language model as the knowledge engine to automatically generate text descriptions for body parts movements of actions, and propose a multi-modal training scheme by utilizing the text encoder to generate feature vectors for different body parts and supervise the skeleton encoder for action representation learning. Experiments show that our proposed GAP method achieves noticeable improvements over various baseline models without extra computation cost at inference. GAP achieves new state-of-the-arts on popular skeleton-based action recognition benchmarks, including NTU RGB+D, NTU RGB+D 120 and NW-UCLA. The source code is available at https://github.com/MartinXM/GAP.
****
Structure Invariant Transformation for better Adversarial Transferability
Xiaosen Wang, Zeliang Zhang, Jianping Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4607-4619
Given the severe vulnerability of Deep Neural Networks (DNNs) against adversarial examples, there is an urgent need for an effective adversarial attack to identify the deficiencies of DNNs in security-sensitive applications. As one of the prevalent black-box adversarial attacks, the existing transfer-based attacks still cannot achieve comparable performance with the white-box attacks. Among these, input transformation based attacks have shown remarkable effectiveness in boosting transferability. In this work, we find that the existing input transformation based attacks transform the input image globally, resulting in limited diversity of the transformed images. We postulate that the more diverse transformed images result in better transferability. Thus, we investigate how to locally apply various transformations onto the input image to improve such diversity while preserving the structure of image. To this end, we propose a novel input transformation based attack, called Structure Invariant Transformation (SIA), which applies a random image transformation onto each image block to craft a set of diverse images for gradient calculation. Extensive experiments on the standard ImageNet dataset demonstrate that SIA exhibits much better transferability than the existing SOTA input transformation based attacks on CNN-based and transformer-based models, showing its generality and superiority in boosting transferability. Code is available at https://github.com/xiaosen-wang/SIT.
****
Thinking Image Color Aesthetics Assessment: Models, Datasets and Benchmarks
Shuai He, Anlong Ming, Yaqi Li, Jinyuan Sun, ShunTian Zheng, Huadong Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21838-21847
We present a comprehensive study on a new task named image color aesthetics assessment (ICAA), which aims to assess color aesthetics based on human perception. ICAA is important for various applications such as imaging measurement and image

analysis. However, due to the highly diverse aesthetic preferences and numerous color combinations, ICAA presents more challenges than conventional image quality assessment tasks.
To advance ICAA research, 1) we propose a baseline model called the Delegate Transformer, which not only deploys deformable transformers to adaptively allocate interest points, but also learns human color space segmentation behavior by the dedicated module.
2) We elaborately build a color-oriented dataset, ICAA17K, containing 17K images, covering 30 popular color combinations, 80 devices and 50 scenes, with each image densely annotated by more than 1,500 people. Moreover, we develop a large-scale benchmark of 15 methods, the most comprehensive one thus far based on two datasets, SPAQ and ICAA17K.
Our work, not only achieves state-of-the-art performance, but more importantly offers the community a roadmap to explore solutions for ICAA. Code and dataset are available in https://github.com/woshidandan/Image-Color-Aesthetics-Assessment.
****
Multi-body Depth and Camera Pose Estimation from Multiple Views
Andrea Porfiri Dal Cin, Giacomo Boracchi, Luca Magri; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17804-17814
Traditional and deep Structure-from-Motion (SfM) methods typically operate under the assumption that the scene is rigid, i.e., the environment is static or consists of a single moving object. Few multi-body SfM approaches address the reconstruction of multiple rigid bodies in a scene but suffer from the inherent scale ambiguity of SfM, such that objects are reconstructed at inconsistent scales. We propose a depth and camera pose estimation framework to resolve the scale ambiguity in multi-body scenes. Specifically, starting from disorganized images, we present a novel multi-view scale estimator that resolves the camera pose ambiguity and a multi-body plane sweep network that generalizes depth estimation to dynamic scenes. Experiments demonstrate the advantages of our method over state-of-the-art SfM frameworks in multi-body scenes and show that it achieves comparable results in static scenes. The code and dataset are available at https://github.com/andreadalcin/MultiBodySfM.
****
DISeR: Designing Imaging Systems with Reinforcement Learning
Tzofi Klinghoffer, Kushagra Tiwary, Nikhil Behari, Bhavya Agrawalla, Ramesh Raskar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23632-23642
Imaging systems consist of cameras to encode visual information about the world and perception models to interpret this encoding. Cameras contain (1) illumination sources, (2) optical elements, and (3) sensors, while perception models use (4) algorithms. Directly searching over all combinations of these four building blocks to design an imaging system is challenging due to the size of the search space. Moreover, cameras and perception models are often designed independently, leading to sub-optimal task performance. In this paper, we formulate these four building blocks of imaging systems as a context-free grammar (CFG), which can be automatically searched over with a learned camera designer to jointly optimize the imaging system with task-specific perception models. By transforming the CFG to a state-action space, we then show how the camera designer can be implemented with reinforcement learning to intelligently search over the combinatorial space of possible imaging system configurations. We demonstrate our approach on two tasks, depth estimation and camera rig design for autonomous vehicles, showing that our method yields rigs that outperform industry-wide standards. We believe that our proposed approach is an important step towards automating imaging system design.
****
The Euclidean Space is Evil: Hyperbolic Attribute Editing for Few-shot Image Generation
Lingxiao Li, Yi Zhang, Shuhui Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22714-22724

Few-shot image generation is a challenging task since it aims to generate diverse new images for an unseen category with only a few images. Existing methods suffer from the trade-off between the quality and diversity of generated images. To tackle this problem, we propose Hyperbolic Attribute Editing (HAE), a simple yet effective method. Unlike prior arts that work in Euclidean space, HAE captures the hierarchy among images using data from seen categories in hyperbolic space. Given a well-trained HAE, images of unseen categories can be generated by moving the latent code of a given image toward any meaningful directions in the Poincare disk with a fixing radius. Most importantly, the hyperbolic space allows us to control the semantic diversity of the generated images by setting different radii in the disk. Extensive experiments and visualizations demonstrate that HAE is capable of not only generating images with promising quality and diversity using limited data but achieving a highly controllable and interpretable editing process.

****

FULLER: Unified Multi-modality Multi-task 3D Perception via Multi-level Gradient Calibration

Zhijian Huang, Sihao Lin, Guiyu Liu, Mukun Luo, Chaoqiang Ye, Hang Xu, Xiaojun Chang, Xiaodan Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3502-3511

Multi-modality fusion and multi-task learning are becoming trendy in 3D autonomous driving scenario, considering robust prediction and computation budget. However, naively extending the existing framework to the domain of multi-modality multi-task learning remains ineffective and even poisonous due to the notorious modality bias and task conflict. Previous works manually coordinate the learning framework with empirical knowledge, which may lead to sub-optima. To mitigate the issue, we propose a novel yet simple multi-level gradient calibration learning framework across tasks and modalities during optimization. Specifically, the gradients, produced by the task heads and used to update the shared backbone, will be calibrated at the backbone's last layer to alleviate the task conflict. Before the calibrated gradients are further propagated to the modality branches of the backbone, their magnitudes will be calibrated again to the same level, ensuring the downstream tasks pay balanced attention to different modalities. Experiments on large-scale benchmark nuScenes demonstrate the effectiveness of the proposed method, eg, an absolute 14.4% mIoU improvement on map segmentation and 1.4% mAP improvement on 3D detection, advancing the application of 3D autonomous driving in the domain of multi-modality fusion and multi-task learning. We also discuss the links between modalities and tasks.

****

Transparent Shape from a Single View Polarization Image

Mingqi Shao, Chongkun Xia, Zhendong Yang, Junnan Huang, Xueqian Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9277-9286

This paper presents a learning-based method for transparent surface estimation from a single view polarization image. Existing shape from polarization(SfP) methods have the difficulty in estimating transparent shape since the inherent transmission interference heavily reduces the reliability of physics-based prior. To address this challenge, we propose the concept of physics-based prior confidence, which is inspired by the characteristic that the transmission component in the polarization image has more noise than reflection. The confidence is used to determine the contribution of the interfered physics-based prior. Then, we build a network(TransSfP) with multi-branch architecture to avoid the destruction of relationships between different hierarchical inputs. To train and test our method, we construct a dataset for transparent shape from polarization with paired polarization images and ground-truth normal maps. Extensive experiments and comparisons demonstrate the superior accuracy of our method.

****

Invariant Feature Regularization for Fair Face Recognition

Jiali Ma, Zhongqi Yue, Kagaya Tomoyuki, Suzuki Tomoki, Karlekar Jayashree, Sugiri Pranata, Hanwang Zhang; Proceedings of the IEEE/CVF International Conference o

n Computer Vision (ICCV), 2023, pp. 20861-20870
Fair face recognition is all about learning invariant feature that generalizes to unseen faces in any demographic group. Unfortunately, face datasets inevitably capture the imbalanced demographic attributes that are ubiquitous in real-world observations, and the model learns biased feature that generalizes poorly in the minority group. We point out that the bias arises due to the confounding demographic attributes, which mislead the model to capture the spurious demographic-specific feature. The confounding effect can only be removed by causal intervention, which requires the confounder annotations. However, such annotations can be prohibitively expensive due to the diversity of the demographic attributes. To tackle this, we propose to generate diverse data partitions iteratively in an unsupervised fashion. Each data partition acts as a self-annotated confounder, enabling our Invariant Feature Regularization (INV-REG) to deconfound. INV-REG is orthogonal to existing methods, and combining INV-REG with two strong baselines (Arcface and CIFP) leads to new state-of-the-art that improves face recognition on a variety of demographic groups. Code is available at https://github.com/milliema/InvReg.
****
Cross-Domain Product Representation Learning for Rich-Content E-Commerce
Xuehan Bai, Yan Li, Yanhua Cheng, Wenjie Yang, Quan Chen, Han Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5697-5706
The proliferation of short video and live-streaming platforms has revolutionized how consumers engage in online shopping. Instead of browsing product pages, consumers are now turning to rich-content e-commerce, where they can purchase products through dynamic and interactive media like short videos and live streams. This emerging form of online shopping has presented new opportunities for platforms to enhance user engagement and shopping experience. However, it has also introduced technical challenges, as products may be presented differently across various media domains. Therefore, a unified product representation is essential for achieving cross-domain product recognition to ensure an optimal user search experience and effective product recommendations. Despite the urgent industrial need for a unified cross-domain product representation, previous studies have predominantly focused only on product pages without taking into account short videos and live streams. To fill the gap in the rich-content e-commerce area, in this paper, we introduce a large-scale cross-domain poduct recognition dataset, called ROPE. ROPE covers a wide range of product categories and contains over 180,000 products, corresponding to millions of short videos and live streams. It is the first dataset to cover product pages, short videos, and live streams simultaneously, providing the basis for establishing a unified product representation across different media domains. Furthermore, we propose a cross-domain product representation framework, namely COPE, which unifies product representations in different domains through multimodal learning including text and vision. Extensive experiments on downstream tasks like cross-modal retrieval and classification demonstrate the effectiveness of COPE in learning a joint feature space for all product domains.
****
DriveAdapter: Breaking the Coupling Barrier of Perception and Planning in End-to-End Autonomous Driving
Xiaosong Jia, Yulu Gao, Li Chen, Junchi Yan, Patrick Langechuan Liu, Hongyang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7953-7963
End-to-end autonomous driving aims to build a fully differentiable system that takes raw sensor data as inputs and directly outputs the planned trajectory or control signals of the ego vehicle. State-of-the-art methods usually follow the `Teacher-Student' paradigm. The Teacher model uses privileged information (ground-truth states of surrounding agents and map elements) to learn the driving strategy. The student model only has access to raw sensor data and conducts behavior cloning on the data collected by the teacher model. By eliminating the noise of the perception part during planning learning, state-of-the-art works could achiev

e better performance with significantly less data compared to those coupled ones.

However, under the current Teacher-Student paradigm, the student model still needs to learn a planning head from scratch, which could be challenging due to the redundant and noisy nature of raw sensor inputs and the casual confusion issue of behavior cloning. In this work, we aim to explore the possibility of directly adopting the strong teacher model to conduct planning while letting the student model focus more on the perception part. We find that even equipped with a SOTA perception model, directly letting the student model learn the required inputs of the teacher model leads to poor driving performance, which comes from the large distribution gap between predicted privileged inputs and the ground-truth.

To this end, we propose DriveAdapter, which employs adapters with the feature alignment objective function between the student (perception) and teacher (planning) modules. Additionally, since the pure learning-based teacher model itself is imperfect and occasionally breaks safety rules, we propose a method of action-guided feature learning with a mask for those imperfect teacher features to further inject the priors of hand-crafted rules into the learning process. DriveAdapter achieves SOTA performance on multiple closed-loop simulation-based benchmarks of CARLA.
****
General Planar Motion from a Pair of 3D Correspondences
Juan Carlos Dibene, Zhixiang Min, Enrique Dunn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8060-8070
We present a novel 2-point method for estimating the relative pose of a camera undergoing planar motion from 3D data (e.g. from a calibrated stereo setup or an RGB-D sensor). Unlike prior art, our formulation does not assume knowledge of the plane of motion, (e.g. parallelism between the optical axis and motion plane) to resolve the under-constrained nature of SE(3) motion estimation in this context. Instead, we enforce geometric constraints identifying, in closed-form, a unique planar motion solution from an orbital set of geometrically consistent SE(3) motion estimates.
We explore the set of special and degenerate geometric cases arising from our formulation.
Experiments on synthetic data characterize the sensitivity of our estimation framework to measurement noise and different types of observed motion.
We integrate our solver within a RANSAC framework and demonstrate robust operation on standard benchmark sequences of real-world imagery.
Code is available at: https://github.com/jdibenes/gpm.
****
Single Depth-image 3D Reflection Symmetry and Shape Prediction
Zhaoxuan Zhang, Bo Dong, Tong Li, Felix Heide, Pieter Peers, Baocai Yin, Xin Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8896-8906
In this paper, we present Iterative Symmetry Completion Network (ISCNet), a single depth-image shape completion method that exploits reflective symmetry cues to obtain more detailed shapes. The efficacy of single depth-image shape completion methods is often sensitive to the accuracy of the symmetry plane. ISCNet therefore jointly estimates the symmetry plane and shape completion iteratively; more complete shapes contribute to more robust symmetry plane estimates and vice versa. Furthermore, our shape completion method operates in the image domain, enabling more efficient high-resolution, detailed geometry reconstruction. We perform the shape completion from pairs of viewpoints, reflected across the symmetry plane, predicted by a reinforcement learning agent to improve robustness and to simultaneously explicitly leverage symmetry. We demonstrate the effectiveness of ISCNet on a variety of object categories on both synthetic and real-scanned datasets.
****
Local Context-Aware Active Domain Adaptation

Tao Sun, Cheng Lu, Haibin Ling; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 18634-18643
Active Domain Adaptation (ADA) queries the labels of a small number of selected
target samples to help adapting a model from a source domain to a target domain.
 The local context of queried data is important, especially when the domain gap
is large. However, this has not been fully explored by existing ADA works. In th
is paper, we propose a Local context-aware ADA framework, named LADA, to address
 this issue. To select informative target samples, we devise a novel criterion b
ased on the local inconsistency of model predictions. Since the labeling budget
is usually small, fine-tuning model on only queried data can be inefficient. We
progressively augment labeled target data with the confident neighbors in a clas
s-balanced manner. Experiments validate that the proposed criterion chooses more
 informative target samples than existing active selection strategies. Furthermo
re, our full method clearly surpasses recent ADA arts on various benchmarks. Cod
e is available at https://github.com/tsun/LADA.
****
Deep Incubation: Training Large Models by Divide-and-Conquering
Zanlin Ni, Yulin Wang, Jiangwei Yu, Haojun Jiang, Yue Cao, Gao Huang; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
17335-17345
Recent years have witnessed a remarkable success of large deep learning models.
However, training these models is challenging due to high computational costs, p
ainfully slow convergence, and overfitting issues. In this paper, we present Dee
p Incubation, a novel approach that enables the efficient and effective training
 of large models by dividing them into smaller sub-modules which can be trained
separately and assembled seamlessly. A key challenge for implementing this idea
is to ensure the compatibility of the independently trained sub-modules. To addr
ess this issue, we first introduce a global, shared meta model, which is leverag
ed to implicitly link all the modules together, and can be designed as an extrem
ely small network with negligible computational overhead. Then we propose a modu
le incubation algorithm, which trains each sub-module to replace the correspondi
ng component of the meta model and accomplish a given learning task. Despite the
 simplicity, our approach effectively encourages each sub-module to be aware of
its role in the target large model, such that the finally-learned sub-modules ca
n collaborate with each other smoothly after being assembled. Empirically, our m
ethod can outperform end-to-end (E2E) training in well-established training sett
ing and shows transferable performance gain for downstream tasks (e.g., object d
etection and image segmentation on COCO and ADE20K). Our code is available at ht
tps://github.com/LeapLabTHU/Deep-Incubation.
****
Downscaled Representation Matters: Improving Image Rescaling with Collaborative
Downscaled Images
Bingna Xu, Yong Guo, Luoqian Jiang, Mianjie Yu, Jian Chen; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12237-12247
Deep networks have achieved great success in image rescaling (IR) task that seek
s to learn the optimal downscaled representations, i.e., low-resolution (LR) ima
ges, to reconstruct the original high-resolution (HR) images. Compared with supe
r-resolution methods that consider a fixed downscaling scheme, e.g., bicubic, IR
 often achieves significantly better reconstruction performance thanks to the le
arned downscaled representations. This highlights the importance of a good downs
caled representation. Existing IR methods mainly learn the downscaled representa
tion by jointly optimizing the downscaling and upscaling models. Unlike them, we
 seek to improve the downscaled representation through a different and more dire
ct way -- directly optimizing the downscaled image itself instead of the down-/u
pscaling models. Consequently, we propose a Hierarchical Collaborative Downscali
ng (HCD) method that performs gradient descent w.r.t. the reconstruction loss in
 both HR and LR domains to improve the downscaled representations, so as to boos
t IR performance. Extensive experiments show that our HCD significantly improves
 the reconstruction performance both quantitatively and qualitatively. Particula
rly, we improve over popular IR methods by >0.57db PSNR on Set5. Moreover, we al

so highlight the flexibility of our HCD since it can generalize well across diverse image rescaling models. The code is available at https://github.com/xubingna/HCD.
****
Detection Transformer with Stable Matching
Shilong Liu, Tianhe Ren, Jiayu Chen, Zhaoyang Zeng, Hao Zhang, Feng Li, Hongyang Li, Jun Huang, Hang Su, Jun Zhu, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6491-6500
This paper is concerned with the matching stability problem across different decoder layers in DEtection TRansformers (DETR). We point out that the unstable matching in DETR is caused by a multi-optimization path problem, which is highlighted by the one-to-one matching design in DETR. To address this problem, we show that the most important design is to use and only use positional metrics (like IOU) to supervise classification scores of positive examples. Under the principle, we propose two simple yet effective modifications by integrating positional metrics to DETR's classification loss and matching cost, named position-supervised loss and position-modulated cost. We verify our methods on several DETR variants. Our methods show consistent improvements over baselines. By integrating our methods with DINO, we achieve 50.4 and 51.5 AP on the COCO detection benchmark using ResNet-50 backbones under 1x (12 epochs) and 2x (24 epochs) training settings, achieving a new record under the same setting. Our code will be made available.
****
Be Everywhere - Hear Everything (BEE): Audio Scene Reconstruction by Sparse Audio-Visual Samples
Mingfei Chen, Kun Su, Eli Shlizerman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7853-7862
Fully immersive and interactive audio-visual scenes are dynamic such that the listeners and the sound emitters move and interact with each other. Reconstruction of an immersive sound experience, as it happens in the scene, requires detailed reconstruction of the audio perceived by the listener at an arbitrary location. The audio at the listener location is a complex outcome of sound propagation through the scene geometry and interacting with surfaces and also the locations of the emitters and the sounds they emit. Due to these aspects, detailed audio reconstruction requires extensive sampling of audio at any potential listener location. This is usually difficult to implement in realistic real-time dynamic scenes. In this work, we propose to circumvent the need for extensive sensors by leveraging audio and visual samples from only a handful of A/V receivers placed in the scene. In particular, we introduce a novel method and end-to-end integrated rendering pipeline which allows the listener to be everywhere and hear everything (BEE) in a dynamic scene in real-time. BEE reconstructs the audio with two main modules, Joint Audio-Visual Representation, and Integrated Rendering Head. The first module extracts the informative audio-visual features of the scene from sparse A/V reference samples, while the second module integrates the audio samples with learned time-frequency transformations to obtain the target sound. Our experiments indicate that BEE outperforms existing methods by a large margin in terms of quality of sound reconstruction, can generalize to scenes not seen in training and runs in real-time speed.
****
iVS-Net: Learning Human View Synthesis from Internet Videos
Junting Dong, Qi Fang, Tianshuo Yang, Qing Shuai, Chengyu Qiao, Sida Peng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22942-22951
Recent advances in implicit neural representations make it possible to generate free-viewpoint videos of the human from sparse view images. To avoid the expensive training for each person, previous methods adopt the generalizable human model and demonstrate impressive results. However, these methods usually rely on limited multi-view images typically collected in the studio or commercial high-quality 3D scans for training, which heavily prohibits their generalization capability for in-the-wild images. To solve this problem, we propose a new approach to l

earn a generalizable human model from a new source of data, i.e., Internet video
s. These videos capture various human appearances and poses and record the perfo
rmers from abundant viewpoints. To exploit these videos, we present a temporal s
elf-supervised pipeline to enforce the local appearance consistency of each body
 part over different frames of the same video. Once learned, the human model ena
bles creating photorealistic free-viewpoint videos from a single input image. Ex
periments show that our method can generate high-quality view synthesis on in-th
e-wild images while only training on monocular videos.
****
Story Visualization by Online Text Augmentation with Context Memory
Daechul Ahn, Daneul Kim, Gwangmo Song, Seung Hwan Kim, Honglak Lee, Dongyeop Kan
g, Jonghyun Choi; Proceedings of the IEEE/CVF International Conference on Comput
er Vision (ICCV), 2023, pp. 3125-3135
Story visualization (SV) is a challenging text-to-image generation task for the
difficulty of not only rendering visual details from the text descriptions but a
lso encoding a longterm context across multiple sentences. While prior efforts m
ostly focus on generating a semantically relevant image for each sentence, encod
ing a context spread across the given paragraph to generate contextually convinc
ing images (e.g., with a correct character or with a proper background of the sc
ene) remains a challenge. To this end, we propose a novel memory architecture fo
r the Bi-directional Transformer framework with an online text augmentation that
 generates multiple pseudo-descriptions as supplementary supervision during trai
ning for better generalization to the language variation at inference. In extens
ive experiments on the two popular SV benchmarks, i.e., the Pororo-SV and Flints
tones-SV, the proposed method significantly outperforms the state of the arts in
 various metrics including FID, character F1, frame accuracy, BLEU-2/3, and R-pr
ecision with similar or less computational complexity.
****
Attention Discriminant Sampling for Point Clouds
Cheng-Yao Hong, Yu-Ying Chou, Tyng-Luh Liu; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 14429-14440
This paper describes an attention-driven approach to 3-D point cloud sampling. W
e establish our method based on a structure-aware attention discriminant analysi
s that explores geometric and semantic relations embodied among points and their
 clusters. The proposed attention discriminant sampling (ADS) starts by efficien
tly decomposing a given point cloud into clusters to implicitly encode its struc
tural and geometric relatedness among points. By treating each cluster as a stru
ctural component, ADS then draws on evaluating two levels of self-attention: wit
hin-cluster and between-cluster. The former reflects the semantic complexity ent
ailed by the learned features of points within each cluster, while the latter re
veals the semantic similarity between clusters. Driven by structurally preservin
g the point distribution, these two aspects of self-attention help avoid samplin
g redundancy and decide the number of sampled points in each cluster. Extensive
experiments demonstrate that ADS significantly improves classification performan
ce to 95.1% on ModelNet40 and 87.5% on ScanObjectNN and achieves 86.9% mIoU on S
hapeNet Part Segmentation. For scene segmentation, ADS yields 91.1% accuracy on
S3DIS with higher mIoU to the state-of-the-art and 75.6% mIoU on ScanNetV2. Furt
hermore, ADS surpasses the state-of-the-art with 55.0% mAP50 on ScanNetV2 object
 detection.
****
Global Balanced Experts for Federated Long-Tailed Learning
Yaopei Zeng, Lei Liu, Li Liu, Li Shen, Shaoguo Liu, Baoyuan Wu; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4815-4
825
Federated learning (FL) is a prevalent distributed machine learning approach tha
t enables collaborative training of a global model across multiple devices witho
ut sharing local data. However, the presence of long-tailed data can negatively
deteriorate the model's performance in real-world FL applications. Moreover, exi
sting re-balance strategies are less effective for the federated long-tailed iss
ue when directly utilizing local label distribution as the class prior at the cl

ients' side. To this end, we propose a novel Global Balanced Multi-Expert (GBME) framework to optimize a balanced global objective, which does not require additional information beyond the standard FL pipeline. In particular, a proxy is derived from the accumulated gradients uploaded by the clients after local training, and is shared by all clients as the class prior for re-balance training. Such a proxy can also guide the client grouping to train a multi-expert model, where the knowledge from different clients can be aggregated via the ensemble of different experts corresponding to different client groups. To further strengthen the privacy-preserving ability, we present a GBME-p algorithm with a theoretical guarantee to prevent privacy leakage from the proxy. Extensive experiments on long-tailed decentralized datasets demonstrate the effectiveness of GBME and GBME-p, both of which show superior performance to state-of-the-art methods.
****
All4One: Symbiotic Neighbour Contrastive Learning via Self-Attention and Redundancy Reduction
Imanol G. Estepa, Ignacio Sarasua, Bhalaji Nagarajan, Petia Radeva; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16243-16253
Nearest neighbour based methods have proved to be one of the most successful self-supervised learning (SSL) approaches due to their high generalization capabilities. However, their computational efficiency decreases when more than one neighbour is used. In this paper, we propose a novel contrastive SSL approach, which we call All4One, that reduces the distance between neighbour representations using "centroids" created through a self-attention mechanism. We use a Centroid Contrasting objective along with single Neighbour Contrasting and Feature Contrasting objectives. Centroids help in learning contextual information from multiple neighbours whereas the neighbour contrast enables learning representations directly from the neighbours and the feature contrast allows learning representations unique to the features.This combination enables All4One to outperform popular instance discrimination approaches by more than 1% on linear classification evaluation for popular benchmark datasets and obtains state-of-the-art (SoTA) results. Finally, we show that All4One is robust towards embedding dimensionalities and augmentations, surpassing NNCLR and Barlow Twins by more than 5% on low dimensionality and weak augmentation settings.
****
Contrastive Pseudo Learning for Open-World DeepFake Attribution
Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, Lizhuang Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20882-20892
The challenge in sourcing attribution for forgery faces has gained widespread attention due to the rapid development of generative techniques. While many recent works have taken essential steps on GAN-generated faces, more threatening attacks related to identity swapping or expression transferring are still overlooked. And the forgery traces hidden in unknown attacks from the open-world unlabeled faces still remain under-explored. To push the related frontier research, we introduce a new benchmark called Open-World DeepFake Attribution (OW-DFA), which aims to evaluate attribution performance against various types of fake faces under open-world scenarios. Meanwhile, we propose a novel framework named Contrastive Pseudo Learning (CPL) for the OW-DFA task through 1) introducing a Global-Local Voting module to guide the feature alignment of forged faces with different manipulated regions, 2) designing a Confidence-based Soft Pseudo-label strategy to mitigate the pseudo-noise caused by similar methods in unlabeled set. In addition, we extend the CPL framework with a multi-stage paradigm that leverages pre-train technique and iterative learning to further enhance traceability performance. Extensive experiments verify the superiority of our proposed method on the OW-DFA and also demonstrate the interpretability of deepfake attribution task and its impact on improving the security of deepfake detection area.
****
ICL-D3IE: In-Context Learning with Diverse Demonstrations Updating for Document Information Extraction

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, Heng Tao Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19485-19494
Large language models (LLMs), such as GPT-3 and ChatGPT, have demonstrated remarkable results in various natural language processing (NLP) tasks with in-context learning, which involves inference based on a few demonstration examples. Despite their successes in NLP tasks, no investigation has been conducted to assess the ability of LLMs to perform document information extraction (DIE) using in-context learning. Applying LLMs to DIE poses two challenges: the modality and task gap. To this end, we propose a simple but effective in-context learning framework called ICL-D3IE, which enables LLMs to perform DIE with different types of demonstration examples. Specifically, we extract the most difficult and distinct segments from hard training documents as hard demonstrations for benefiting all test instances. We design demonstrations describing relationships that enable LLMs to understand positional relationships. We introduce formatting demonstrations for easy answer extraction. Additionally, the framework improves diverse demonstrations by updating them iteratively. Our experiments on three widely used benchmark datasets demonstrate that the ICL-D3IE framework enables GPT-3/ChatGPT to achieve superior performance when compared to previous pre-trained methods fine-tuned with full training in both the in-distribution (ID) setting and in the out-of-distribution (OOD) setting. Code is available at https://anonymous.4open.science/r/ICL-D3IE-B1EE.
****
IHNet: Iterative Hierarchical Network Guided by High-Resolution Estimated Information for Scene Flow Estimation
Yun Wang, Cheng Chi, Min Lin, Xin Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10073-10082
Scene flow estimation, which predicts the 3D displacements of point clouds, is a fundamental task in autonomous driving. Most methods have adopted a coarse-to-fine structure to balance computational efficiency with accuracy, particularly when handling large displacements. However, inaccuracies in the initial coarse layer's scene flow estimates may accumulate, leading to incorrect final estimates. To alleviate this, we introduce a novel Iterative Hierarchical Network----IHNet. This approach circulates high-resolution estimated information (scene flow and feature) from the preceding iteration back to the low-resolution layer of the current iteration. Serving as a guide, the high-resolution estimated scene flow, instead of initializing the scene flow from zero, provides a more precise center for low-resolution layer to identify matches. Meanwhile, the decoder's feature at the high-resolution layer can contribute essential movement information. Furthermore, based on the recurrent structure, we design a resampling scheme to enhance the correspondence between points across two consecutive frames. By employing the previous estimated scene flow to fine-tune the target frame's coordinates, we can significantly reduce the correspondence discrepancy between two frame points, a problem often caused by point sparsity. Following this adjustment, we continue to estimate the scene flow using the newly updated coordinates, along with the reencoded feature. Our approach outperforms the recent state-of-the-art method WSAFlowNet by 20.1% on FlyingThings3D and 56.0% on KITTI scene flow datasets according to EPE3D metric. The code is available at https://github.com/wangyunlhr/IHNet.
****
SimNP: Learning Self-Similarity Priors Between Neural Points
Christopher Wewer, Eddy Ilg, Bernt Schiele, Jan Eric Lenssen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8841-8852
Existing neural field representations for 3D object reconstruction either (1) utilize object-level representations, but suffer from low-quality details due to conditioning on a global latent code, or (2) are able to perfectly reconstruct the observations, but fail to utilize object-level prior knowledge to infer unobserved regions. We present SimNP, a method to learn category-level self-similarities, which combines the advantages of both worlds by connecting neural point radi

ance fields with a category-level self-similarity representation. Our contribution is two-fold. (1) We design the first neural point representation on a category level by utilizing the concept of coherent point clouds. The resulting neural point radiance fields store a high level of detail for locally supported object regions. (2) We learn how information is shared between neural points in an unconstrained and unsupervised fashion, which allows to derive unobserved regions of an object during the reconstruction process from given observations. We show that SimNP is able to outperform previous methods in reconstructing symmetric unseen object regions, surpassing methods that build upon category-level or pixel-aligned radiance fields, while providing semantic correspondences between instances.
****
Beyond the Limitation of Monocular 3D Detector via Knowledge Distillation
Yiran Yang, Dongshuo Yin, Xuee Rong, Xian Sun, Wenhui Diao, Xinming Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9077-9086
Knowledge distillation (KD) is a promising approach that facilitates the compact student model to learn dark knowledge from the huge teacher model for better results. Although KD methods are well explored in the 2D detection task, existing approaches are not suitable for 3D monocular detection without considering spatial cues. Motivated by the potential of depth information, we propose a novel distillation framework that validly improves the performance of the student model without extra depth labels. Specifically, we first put forward a perspective-induced feature imitation, which utilizes the perspective principle (the farther the smaller) to facilitate the student to imitate more features of farther objects from the teacher model. Moreover, we construct a depth-guided matrix by the predicted depth gap of teacher and student to facilitate the model to learn more knowledge of farther objects in prediction level distillation. The proposed method is available for advanced monocular detectors with various backbones, which also brings no extra inference time. Extensive experiments on the KITTI and nuScenes benchmarks with diverse settings demonstrate that the proposed method outperforms the state-of-the-art KD methods.
****
Cascade-DETR: Delving into High-Quality Universal Object Detection
Mingqiao Ye, Lei Ke, Siyuan Li, Yu-Wing Tai, Chi-Keung Tang, Martin Danelljan, Fisher Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6704-6714
Object localization in general environments is a fundamental part of vision systems. While dominating on the COCO benchmark, recent Transformer-based detection methods are not competitive in diverse domains. Moreover, these methods still struggle to very accurately estimate the object bounding boxes in complex environments.

We introduce Cascade-DETR for high-quality universal object detection. We jointly tackle the generalization to diverse domains and localization accuracy by proposing the Cascade Attention layer, which explicitly integrates object-centric information into the detection decoder by limiting the attention to the previous box prediction. To further enhance accuracy, we also revisit the scoring of queries. Instead of relying on classification scores, we predict the expected IoU of the query, leading to substantially more well-calibrated confidences. Lastly, we introduce a universal object detection benchmark, UDB10, that contains 10 datasets from diverse domains. While also advancing the state-of-the-art on COCO, Cascade-DETR substantially improves DETR-based detectors on all datasets in UDB10, even by over 10 mAP in some cases. The improvements under stringent quality requirements are even more pronounced. Our code and pretrained models are at https://github.com/SysCV/cascade-detr.
****
ACLS: Adaptive and Conditional Label Smoothing for Network Calibration
Hyekang Park, Jongyoun Noh, Youngmin Oh, Donghyeon Baek, Bumsub Ham; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3

936-3945
We address the problem of network calibration adjusting miscalibrated confidence
s of deep neural networks. Many approaches to network calibration adopt a regula
rization-based method that exploits a regularization term to smooth the miscalib
rated confidences. Although these approaches have shown the effectiveness on cal
ibrating the networks, there is still a lack of understanding on the underlying
principles of regularization in terms of network calibration. We present in this
 paper an in-depth analysis of existing regularization-based methods, providing
a better understanding on how they affect to network calibration. Specifically,
we have observed that 1) the regularization-based methods can be interpreted as
variants of label smoothing, and 2) they do not always behave desirably. Based o
n the analysis, we introduce a novel loss function, dubbed ACLS, that unifies th
e merits of existing regularization methods, while avoiding the limitations. We
show extensive experimental results for image classification and semantic segmen
tation on standard benchmarks, including CIFAR10, Tiny-ImageNet, ImageNet, and P
ASCAL VOC, demonstrating the effectiveness of our loss function.
****
EMR-MSF: Self-Supervised Recurrent Monocular Scene Flow Exploiting Ego-Motion Ri
gidity
Zijie Jiang, Masatoshi Okutomi; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 69-78
Self-supervised monocular scene flow estimation, aiming to understand both 3D st
ructures and 3D motions from two temporally consecutive monocular images, has re
ceived increasing attention for its simple and economical sensor setup. However,
 the accuracy of current methods suffers from the bottleneck of less-efficient n
etwork architecture and lack of motion rigidity for regularization. In this pape
r, we propose a superior model named EMR-MSF by borrowing the advantages of netw
ork architecture design under the scope of supervised learning. We further impos
e explicit and robust geometric constraints with an elaborately constructed ego-
motion aggregation module where a rigidity soft mask is proposed to filter out d
ynamic regions for stable ego-motion estimation using static regions. Moreover,
we propose a motion consistency loss along with a mask regularization loss to fu
lly exploit static regions. Several efficient training strategies are integrated
 including a gradient detachment technique and an enhanced view synthesis proces
s for better performance. Our proposed method outperforms the previous self-supe
rvised works by a large margin and catches up to the performance of supervised m
ethods. On the KITTI scene flow benchmark, our approach improves the SF-all metr
ic of the state-of-the-art self-supervised monocular method by 44% and demonstra
tes superior performance across sub-tasks including depth and visual odometry, a
mongst other self-supervised single-task or multi-task methods.
****
Evaluation and Improvement of Interpretability for Self-Explainable Part-Prototy
pe Networks
Qihan Huang, Mengqi Xue, Wenqi Huang, Haofei Zhang, Jie Song, Yongcheng Jing, Mi
ngli Song; Proceedings of the IEEE/CVF International Conference on Computer Visi
on (ICCV), 2023, pp. 2011-2020
Part-prototype networks (e.g., ProtoPNet, ProtoTree, and ProtoPool) have attract
ed broad research interest for their intrinsic interpretability and comparable a
ccuracy to non-interpretable counterparts. However, recent works find that the i
nterpretability from prototypes is fragile, due to the semantic gap between the
similarities in the feature space and that in the input space. In this work, we
strive to address this challenge by making the first attempt to quantitatively a
nd objectively evaluate the interpretability of the part-prototype networks. Spe
cifically, we propose two evaluation metrics, termed as "consistency score" and
"stability score", to evaluate the explanation consistency across images and the
 explanation robustness against perturbations, respectively, both of which are e
ssential for explanations taken into practice. Furthermore, we propose an elabor
ated part-prototype network with a shallow-deep feature alignment (SDFA) module
and a score aggregation (SA) module to improve the interpretability of prototype
s. We conduct systematical evaluation experiments and provide substantial discus

sions to uncover the interpretability of existing part-prototype networks. Exper iments on three benchmarks across nine architectures demonstrate that our model achieves significantly superior performance to the state of the art, in both the accuracy and interpretability. Our code is available at https://github.com/hqhQ AQ/EvalProtoPNet.
****
Temporal-Coded Spiking Neural Networks with Dynamic Firing Threshold: Learning w ith Event-Driven Backpropagation
Wenjie Wei, Malu Zhang, Hong Qu, Ammar Belatreche, Jian Zhang, Hong Chen; Procee dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10552-10562
Spiking Neural Networks (SNNs) offer a highly promising computing paradigm due t o their biological plausibility, exceptional spatiotemporal information processi ng capability and low power consumption. As a temporal encoding scheme for SNNs, Time-To-First-Spike (TTFS) encodes information using the timing of a single spi ke, which allows spiking neurons to transmit information through sparse spike tr ains and results in lower power consumption and higher computational efficiency compared to traditional rate-based encoding counterparts. However, despite the a dvantages of the TTFS encoding scheme, the effective and efficient training of T TFS-based deep SNNs remains a significant and open research problem. In this wor k, we first examine the factors underlying the limitations of applying existing TTFS-based learning algorithms to deep SNNs. Specifically, we investigate issues related to over-sparsity of spikes and the complexity of finding the `causal se t'. We then propose a simple yet efficient dynamic firing threshold (DFT) mechan ism for spiking neurons to address these issues. Building upon the proposed DFT mechanism, we further introduce a novel direct training algorithm for TTFS-based deep SNNs, called DTA-TTFS. This method utilizes event-driven processing and sp ike timing to enable efficient learning of deep SNNs. The proposed training meth od was validated on the image classification task and experimental results clear ly demonstrate that our proposed method achieves state-of-the-art accuracy in co mparison to existing TTFS-based learning algorithms, while maintaining high leve ls of sparsity and energy efficiency on neuromorphic inference accelerator.
****
Mitigating Adversarial Vulnerability through Causal Parameter Estimation by Adve rsarial Double Machine Learning
Byung-Kwan Lee, Junho Kim, Yong Man Ro; Proceedings of the IEEE/CVF Internationa l Conference on Computer Vision (ICCV), 2023, pp. 4499-4509
Adversarial examples derived from deliberately crafted perturbations on visual i nputs can easily harm decision process of deep neural networks. To prevent poten tial threats, various adversarial training-based defense methods have grown rapi dly and become a de facto standard approach for robustness. Despite recent compe titive achievements, we observe that adversarial vulnerability varies across tar gets and certain vulnerabilities remain prevalent. Intriguingly, such peculiar p henomenon cannot be relieved even with deeper architectures and advanced defense methods. To address this issue, in this paper, we introduce a causal approach c alled Adversarial Double Machine Learning (ADML), which allows us to quantify th e degree of adversarial vulnerability for network predictions and capture the ef fect of treatments on outcome of interests. ADML can directly estimate causal pa rameter of adversarial perturbations per se and mitigate negative effects that c an potentially damage robustness, bridging a causal perspective into the adversa rial vulnerability. Through extensive experiments on various CNN and Transformer architectures, we corroborate that ADML improves adversarial robustness with la rge margins and relieve the empirical observation.
****
Dynamic Token Pruning in Plain Vision Transformers for Semantic Segmentation
Quan Tang, Bowen Zhang, Jiajun Liu, Fagui Liu, Yifan Liu; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 777-786
Vision transformers have achieved leading performance on various visual tasks ye t still suffer from high computational complexity. The situation deteriorates in dense prediction tasks like semantic segmentation, as high-resolution inputs an

d outputs usually imply more tokens involved in computations. Directly removing
the less attentive tokens has been discussed for the image classification task b
ut can not be extended to semantic segmentation since a dense prediction is requ
ired for every patch. To this end, this work introduces a Dynamic Token Pruning
(DToP) method based on the early exit of tokens for semantic segmentation. Motiv
ated by the coarse-to-fine segmentation process by humans, we naturally split th
e widely adopted auxiliary-loss-based network architecture into several stages,
where each auxiliary block grades every token's difficulty level. We can finaliz
e the prediction of easy tokens in advance without completing the entire forward
 pass. Moreover, we keep k highest confidence tokens for each semantic category
to uphold the representative context information. Thus, computational complexity
 will change with the difficulty of the input, akin to the way humans do segment
ation. Experiments suggest that the proposed DToP architecture reduces on averag
e 20% 35% of computational cost for current semantic segmentation methods based
on plain vision transformers without accuracy degradation. The code is available
 through the following link: https://github.com/zbwxp/Dynamic-Token-Pruning.
****
Shape Anchor Guided Holistic Indoor Scene Understanding
Mingyue Dong, Linxi Huan, Hanjiang Xiong, Shuhan Shen, Xianwei Zheng; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
21916-21926
This paper proposes a shape anchor guided learning strategy (AncLearn) for robus
t holistic indoor scene understanding. We observe that the search space construc
ted by current methods for proposal feature grouping and instance point sampling
 often introduces massive noise to instance detection and mesh reconstruction. A
ccordingly, we develop AncLearn to generate anchors that dynamically fit instanc
e surfaces to (i) unmix noise and target-related features for offering reliable
proposals at the detection stage, and (ii) reduce outliers in object point sampl
ing for directly providing well-structured geometry priors without segmentation
during reconstruction. We embed AncLearn into a reconstruction-from-detection le
arning system (AncRec) to generate high-quality semantic scene models in a purel
y instance-oriented manner. Experiments conducted on the ScanNetv2 dataset (with
 ground truths from Scan2CAD and SceneCAD) demonstrate that our shape anchor-bas
ed method consistently achieves state-of-the-art performance in terms of 3D obje
ct detection, layout estimation, and shape reconstruction.
****
Knowledge-Aware Federated Active Learning with Non-IID Data
Yu-Tong Cao, Ye Shi, Baosheng Yu, Jingya Wang, Dacheng Tao; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22279-2228
9
Federated learning enables multiple decentralized clients to learn collaborative
ly without sharing local data. However, the expensive annotation cost on local c
lients remains an obstacle in utilizing local data. In this paper, we propose a
federated active learning paradigm to efficiently learn a global model with a li
mited annotation budget while protecting data privacy in a decentralized learnin
g manner. The main challenge faced by federated active learning is the mismatch
between the active sampling goal of the global model on the server and that of t
he asynchronous local clients. This becomes even more significant when data is d
istributed non-IID across local clients. To address the aforementioned challenge
, we propose Knowledge-Aware Federated Active Learning (KAFAL), which consists o
f Knowledge-Specialized Active Sampling (KSAS) and Knowledge-Compensatory Federa
ted Update (KCFU). Specifically, KSAS is a novel active sampling method tailored
 for the federated active learning problem, aiming to deal with the mismatch cha
llenge by sampling actively based on the discrepancies between local and global
models. KSAS intensifies specialized knowledge in local clients, ensuring the sa
mpled data is informative for both the local clients and the global model. Meanw
hile, KCFU deals with the client heterogeneity caused by limited data and non-II
D data distributions by compensating for each client's ability in weak classes w
ith the assistance of the global model. Extensive experiments and analyses are c
onducted to show the superiority of KAFAL over recent state-of-the-art active le

arning methods. Code is available at https://github.com/ycao5602/KAFAL.
****
PlankAssembly: Robust 3D Reconstruction from Three Orthographic Views with Learnt Shape Programs
Wentao Hu, Jia Zheng, Zixin Zhang, Xiaojun Yuan, Jian Yin, Zihan Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18495-18505
In this paper, we develop a new method to automatically convert 2D line drawings from three orthographic views into 3D CAD models. Existing methods for this problem reconstruct 3D models by back-projecting the 2D observations into 3D space while maintaining explicit correspondence between the input and output. Such methods are sensitive to errors and noises in the input, thus often fail in practice where the input drawings created by human designers are imperfect. To overcome this difficulty, we leverage the attention mechanism in a Transformer-based sequence generation model to learn flexible mappings between the input and output. Further, we design shape programs which are suitable for generating the objects of interest to boost the reconstruction accuracy and facilitate CAD modeling applications. Experiments on a new benchmark dataset show that our method significantly outperforms existing ones when the inputs are noisy or incomplete.
****
PODIA-3D: Domain Adaptation of 3D Generative Model Across Large Domain Gap Using Pose-Preserved Text-to-Image Diffusion
Gwanghyun Kim, Ji Ha Jang, Se Young Chun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22603-22612
Recently, significant advancements have been made in 3D generative models, however training these models across diverse domains is challenging and requires an huge amount of training data and knowledge of pose distribution. Text-guided domain adaptation methods have allowed the generator to be adapted to the target domains using text prompts, thereby obviating the need for assembling numerous data. Recently, DATID-3D presents impressive quality of samples in text-guided domain, preserving diversity in text by leveraging text-to-image diffusion. However, adapting 3D generators to domains with significant domain gaps from the source domain still remains challenging due to issues in current text-to-image diffusion models as following: 1) shape-pose trade-off in diffusion-based translation, 2) pose bias, and 3) instance bias in the target domain, resulting in inferior 3D shapes, low text-image correspondence, and low intra-domain diversity in the generated samples. To address these issues, we propose a novel pipeline called PODIA-3D, which uses pose-preserved text-to-image diffusion-based domain adaptation for 3D generative models. We construct a pose-preserved text-to-image diffusion model that allows the use of extremely high-level noise for significant domain changes. We also propose specialized-to-general sampling strategies to improve the details of the generated samples. Moreover, to overcome the instance bias, we introduce a text-guided debiasing method that improves intra-domain diversity. Consequently, our method successfully adapts 3D generators across significant domain gaps. Our qualitative results and user study demonstrate that our approach outperforms existing 3D text-guided domain adaptation methods in terms of text-image correspondence, realism, diversity of rendered images, and sense of depth of 3D shapes in the generated samples.
****
Steered Diffusion: A Generalized Framework for Plug-and-Play Conditional Image Synthesis
Nithin Gopalakrishnan Nair, Anoop Cherian, Suhas Lohit, Ye Wang, Toshiaki Koike-Akino, Vishal M. Patel, Tim K. Marks; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20850-20860
Conditional generative models typically demand large annotated training sets to achieve high-quality synthesis. As a result, there has been significant interest in designing models that perform plug-and-play generation, i.e., to use a predefined or pretrained model, which is not explicitly trained on the generative task, to guide the generative process (e.g., using language). However, such guidance is typically useful only towards synthesizing high-level semantics rather than

editing fine-grained details as in image-to-image translation tasks. To this end, and capitalizing on the powerful fine-grained generative control offered by the recent diffusion-based generative models, we introduce Steered Diffusion, a generalized framework for photorealistic zero-shot conditional image generation using a diffusion model trained for unconditional generation. The key idea is to steer the image generation of the diffusion model at inference time via designing a loss using a pre-trained inverse model that characterizes the conditional task. This loss modulates the sampling trajectory of the diffusion process. Our framework allows for easy incorporation of multiple conditions during inference. We present experiments using steered diffusion on several tasks including inpainting, colorization, text-guided semantic editing, and image super-resolution. Our results demonstrate clear qualitative and quantitative improvements over state-of-the-art diffusion-based plug-and-play models while adding negligible additional computational cost.
****
DiffFit: Unlocking Transferability of Large Diffusion Models via Simple Parameter-efficient Fine-Tuning
Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, Zhenguo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4230-4239
Diffusion models have proven to be highly effective in generating high-quality images. However, adapting large pre-trained diffusion models to new domains remains an open challenge, which is critical for real-world applications. This paper proposes DiffFit, a parameter-efficient strategy to fine-tune large pre-trained diffusion models that enable fast adaptation to new domains. DiffFit is embarrassingly simple that only fine-tunes the bias term and newly-added scaling factors in specific layers, yet resulting in significant training speed-up and reduced model storage costs. Compared with full fine-tuning, DiffFit achieves 2x training speed-up and only needs to store approximately 0.12% of the total model parameters. Intuitive theoretical analysis has been provided to justify the efficacy of scaling factors on fast adaptation. On 8 downstream datasets, DiffFit achieves superior or competitive performances compared to the full fine-tuning while being more efficient. Remarkably, we show that DiffFit can adapt a pre-trained low-resolution generative model to a high-resolution one by adding minimal cost. Among diffusion-based methods, DiffFit sets a new state-of-the-art FID of 3.02 on ImageNet 512x512 benchmark by fine-tuning only 25 epochs from a public pre-trained ImageNet 256x256 checkpoint while being 30x more training efficient than the closest competitor.
****
NeO 360: Neural Fields for Sparse View Synthesis of Outdoor Scenes
Muhammad Zubair Irshad, Sergey Zakharov, Katherine Liu, Vitor Guizilini, Thomas Kollar, Adrien Gaidon, Zsolt Kira, Rares Ambrus; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9187-9198
Recent implicit neural representations have shown great results for novel view synthesis. However, existing methods require expensive per-scene optimization from many views hence limiting their application to real-world unbounded urban settings where the objects of interest or backgrounds are observed from very few views. To mitigate this challenge, we introduce a new approach called NeO 360, Neural fields for sparse view synthesis of outdoor scenes. NeO 360 is a generalizable method that reconstructs 360deg scenes from a single or a few posed RGB images. The essence of our approach is in capturing the distribution of complex real-world outdoor 3D scenes and using a hybrid image-conditional triplanar representation that can be queried from any world point. Our representation combines the best of both voxel-based and bird's-eye-view (BEV) representations and is more effective and expressive than each. NeO 360's representation allows us to learn from a large collection of unbounded 3D scenes while offering generalizability to new views and novel scenes from as few as a single image during inference. We demonstrate our approach on the proposed challenging 360deg unbounded dataset, called NeRDS 360, and show that NeO 360 outperforms state-of-the-art generalizable methods for novel view synthesis while also offering editing and composition cap

abilities. Project page: zubair-irshad.github.io/projects/neo360.html
****
UnLoc: A Unified Framework for Video Localization Tasks
Shen Yan, Xuehan Xiong, Arsha Nagrani, Anurag Arnab, Zhonghao Wang, Weina Ge, David Ross, Cordelia Schmid; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13623-13633
While large-scale image-text pretrained models such as CLIP have been used for multiple video-level tasks on trimmed videos, their use for temporal localization in untrimmed videos is still a relatively unexplored task. We design a new approach for this called UnLoc, which uses pretrained image and text towers, and feeds tokens to a video-text fusion model. The output of the fusion module are then used to construct a feature pyramid in which each level connects to a head to predict a per-frame relevancy score and start/end time displacements. Unlike previous works, our architecture enables Moment Retrieval, Temporal Localization, and Action Segmentation with a single stage model, without the need for action proposals, motion based pretrained features or representation masking. Unlike specialized models, we achieve state of the art results on all three different localization tasks with a unified approach. Code is available at: https://github.com/google-research/scenic.
****
QD-BEV : Quantization-aware View-guided Distillation for Multi-view 3D Object Detection
Yifan Zhang, Zhen Dong, Huanrui Yang, Ming Lu, Cheng-Ching Tseng, Yuan Du, Kurt Keutzer, Li Du, Shanghang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3825-3835
Multi-view 3D detection based on BEV (bird-eye-view) has recently achieved significant improvements. However, the huge memory consumption of state-of-the-art models makes it hard to deploy them on vehicles, and the non-trivial latency will affect the real-time perception of streaming applications. Despite the wide application of quantization to lighten models, we show in our paper that directly applying quantization in BEV tasks will 1) make the training unstable, and 2) lead to intolerable performance degradation. To solve these issues, our method QD-BEV enables a novel view-guided distillation (VGD) objective, which can stabilize the quantization-aware training (QAT) while enhancing the model performance by leveraging both image features and BEV features. Our experiments show that QD-BEV achieves similar or even better accuracy than previous methods with significant efficiency gains. On the nuScenes datasets, the 4-bit weight and 6-bit activation quantized QD-BEV-Tiny model achieves 37.2% NDS with only 15.8 MB model size, outperforming BevFormer-Tiny by 1.8% with an 8x model compression. On the Small and Base variants, QD-BEV models also perform superbly and achieve 47.9% NDS (28.2 MB) and 50.9% NDS (32.9 MB), respectively.
****
Fast Inference and Update of Probabilistic Density Estimation on Trajectory Prediction
Takahiro Maeda, Norimichi Ukita; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9795-9805
Safety-critical applications such as autonomous vehicles and social robots require fast computation and accurate probability density estimation on trajectory prediction. To address both requirements, this paper presents a new normalizing flow-based trajectory prediction model named FlowChain. FlowChain is a stack of conditional continuously-indexed flows (CIFs) that are expressive and allow analytical probability density computation. This analytical computation is faster than the generative models that need additional approximations such as kernel density estimation. Moreover, FlowChain is more accurate than the Gaussian mixture-based models due to fewer assumptions on the estimated density. FlowChain also allows a rapid update of estimated probability densities. This update is achieved by adopting the newest observed position and reusing the flow transformations and its log-det-jacobians that represent the motion trend. This update is completed in less than one millisecond because this reuse greatly omits the computational cost. Experimental results showed our FlowChain achieved state-of-the-art trajec

tory prediction accuracy compared to previous methods. Furthermore, our FlowChai
n demonstrated superiority in the accuracy and speed of density estimation. Our
code is available at https://github.com/meaten/FlowChain-ICCV2023
****
CLIPascene: Scene Sketching with Different Types and Levels of Abstraction
Yael Vinker, Yuval Alaluf, Daniel Cohen-Or, Ariel Shamir; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4146-4156
In this paper, we present a method for converting a given scene image into a ske
tch using different types and multiple levels of abstraction. We distinguish bet
ween two types of abstraction.
 The first considers the fidelity of the sketch, varying its representation from
 a more precise portrayal of the input to a looser depiction.
 The second is defined by the visual simplicity of the sketch, moving from a det
ailed depiction to a sparse sketch.
 Using an explicit disentanglement into two abstraction axes --- and multiple le
vels for each one --- provides users additional control over selecting the desir
ed sketch based on their personal goals and preferences.
 To form a sketch at a given level of fidelity and simplification, we train two
MLP networks. The first network learns the desired placement of strokes, while t
he second network learns to gradually remove strokes from the sketch without har
ming its recognizability and semantics.
 Our approach is able to generate sketches of complex scenes including those wit
h complex backgrounds (e.g. natural and urban settings) and subjects (e.g. anima
ls and people) while depicting gradual abstractions of the input scene in terms
of fidelity and simplicity.
****
Vision Grid Transformer for Document Layout Analysis
Cheng Da, Chuwei Luo, Qi Zheng, Cong Yao; Proceedings of the IEEE/CVF Internatio
nal Conference on Computer Vision (ICCV), 2023, pp. 19462-19472
Document pre-trained models and grid-based models have proven to be very effecti
ve on various tasks in Document AI. However, for the document layout analysis (D
LA) task, existing document pre-trained models, even those pre-trained in a mult
i-modal fashion, usually rely on either textual features or visual features. Gri
d-based models for DLA are multi-modality but largely neglect the effect of pre-
training. To fully leverage multi-modal information and exploit pre-training tec
hniques to learn better representation for DLA, in this paper, we present VGT, a
 two-stream Vision Grid Transformer, in which Grid Transformer (GiT) is proposed
 and pre-trained for 2D token-level and segment-level semantic understanding. Fu
rthermore, a new dataset named D^4LA, which is so far the most diverse and detai
led manually-annotated benchmark for document layout analysis, is curated and re
leased. Experiment results have illustrated that the proposed VGT model achieves
 new state-of-the-art results on DLA tasks, e.g. PubLayNet (95.7% to 96.2%), Doc
Bank (79.6% to 84.1%), and D^4LA (67.7% to 68.8%). The code and models as well a
s the D4LA dataset will be made publicly available.
****
Multi-Directional Subspace Editing in Style-Space
Chen Naveh; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2023, pp. 7138-7148
This paper describes a new technique for finding disentangled semantic direction
s in the latent space of StyleGAN. Our method identifies meaningful orthogonal s
ubspaces that allow editing of one human face attribute, while minimizing undesi
red changes in other attributes. Our model is capable of editing a single attrib
ute in multiple directions, resulting in a range of possible generated images. W
e compare our scheme with three state-of-the-art models and show that our method
 outperforms them in terms of face editing and disentanglement capabilities. Add
itionally, we suggest quantitative measures for evaluating attribute separation
and disentanglement, and exhibit the superiority of our model with respect to th
ose measures.
****
Adaptive Superpixel for Active Learning in Semantic Segmentation

Hoyoung Kim, Minhyeon Oh, Sehyun Hwang, Suha Kwak, Jungseul Ok; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 943-953

Learning semantic segmentation requires pixel-wise annotations, which can be time-consuming and expensive. To reduce the annotation cost, we propose a superpixel-based active learning (AL) framework, which collects a dominant label per superpixel instead. To be specific, it consists of adaptive superpixel and sieving mechanisms, fully dedicated to AL. At each round of AL, we adaptively merge neighboring pixels of similar learned features into superpixels. We then query a selected subset of these superpixels using an acquisition function assuming no uniform superpixel size. This approach is more efficient than existing methods, which rely only on innate features such as RGB color and assume uniform superpixel sizes. Obtaining a dominant label per superpixel drastically reduces annotators' burden as it requires fewer clicks. However, it inevitably introduces noisy annotations due to mismatches between superpixel and ground truth segmentation. To address this issue, we further devise a sieving mechanism that identifies and excludes potentially noisy annotations from learning. Our experiments on both Cityscapes and PASCAL VOC datasets demonstrate the efficacy of adaptive superpixel and sieving mechanisms.

****

Adaptive Spiral Layers for Efficient 3D Representation Learning on Meshes

Francesca Babiloni, Matteo Maggioni, Thomas Tanay, Jiankang Deng, Ales Leonardis, Stefanos Zafeiriou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14620-14631

The success of deep learning models on structured data has generated significant interest in extending their application to non-Euclidean domains. In this work, we introduce a novel intrinsic operator suitable for representation learning on 3D meshes. Our operator is specifically tailored to adapt its behavior to the irregular structure of the underlying graph and effectively utilize its long-range dependencies, while at the same time ensuring computational efficiency and ease of optimization. In particular, inspired by the framework of Spiral Convolution, which extracts and transforms the vertices in the 3D mesh following a local spiral ordering, we propose a general operator that dynamically adjusts the length of the spiral trajectory and the parameters of the transformation for each processed vertex and mesh. Then, we use polyadic decomposition to factorize its dense weight tensor into a sequence of lighter linear layers that separately process features and vertices information, hence significantly reducing the computational complexity without introducing any stringent inductive biases. Notably, we leverage dynamic gating to achieve spatial adaptivity and induce global reasoning with constant time complexity benefitting from an efficient dynamic pooling mechanism based on Summed-Area-tables. Used as a drop-in replacement on existing architectures for shape correspondence our operator significantly improves the performance-efficiency trade-off, and in 3D shape generation with morphable models achieves state-of-the-art performance with a three-fold reduction in the number of parameters required. Project page: https://github.com/Fb2221/DFC

****

Parametric Information Maximization for Generalized Category Discovery

Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, Ismail Ben Ayed; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1729-1739

We introduce a Parametric Information Maximization (PIM) model for the Generalized Category Discovery (GCD) problem. Specifically, we propose a bi-level optimization formulation, which explores a parameterized family of objective functions, each evaluating a weighted mutual information between the features and the latent labels, subject to supervision constraints from the labeled samples. Our formulation mitigates the class-balance bias encoded in standard information maximization approaches, thereby handling effectively both short-tailed and long-tailed data sets. We report extensive experiments and comparisons demonstrating that our PIM model consistently sets new state-of-the-art performances in GCD across six different datasets, more so when dealing with challenging fine-grained proble

ms. Our code: https://github.com/ThalesGroup/pim-generalized-category-discovery.
****
Convex Decomposition of Indoor Scenes
Vaibhav Vavilala, David Forsyth; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2023, pp. 9176-9186
We describe a method to parse a complex, cluttered indoor scene into primitives which offer a parsimonious abstraction of scene structure. Our primitives are si mple convexes. Our method uses a learned regression procedure to parse a scene i nto a fixed number of convexes from RGBD input, and can optionally accept segmen tations to improve the decomposition. The result is then polished with a descent method which adjusts the convexes to produce a very good fit, and greedily remo ves superfluous primitives. Because the entire scene is parsed, we can evaluate using traditional depth, normal, and segmentation error metrics. Our evaluation procedure demonstrates that the error from our primitive representation is compa rable to that of predicting depth from a single image.
****
Toward Unsupervised Realistic Visual Question Answering
Yuwei Zhang, Chih-Hui Ho, Nuno Vasconcelos; Proceedings of the IEEE/CVF Internat ional Conference on Computer Vision (ICCV), 2023, pp. 15613-15624
The problem of realistic VQA (RVQA), where a model has to reject unanswerable qu estions (UQs) and answer answerable ones (AQs), is studied. We first point out 2 drawbacks in current RVQA research, where (1) datasets contain too many unchall enging UQs and (2) a large number of annotated UQs are required for training. To resolve the first drawback, we propose a new testing dataset, RGQA, which combi nes AQs from an existing VQA dataset with around 29K human-annotated UQs. These UQs consist of both fine-grained and coarse-grained image-question pairs generat ed with 2 approaches: CLIP-based and Perturbation-based. To address the second d rawback, we introduce an unsupervised training approach. This combines pseudo UQ s obtained by randomly pairing images and questions, with an RoI Mixup procedure to generate more fine-grained pseudo UQs, and model ensembling to regularize mo del confidence. Experiments show that using pseudo UQs significantly outperforms RVQA baselines. RoI Mixup and model ensembling further increase the gain. Final ly, human evaluation reveals a performance gap between humans and models, showin g that more RVQA research is needed.
****
A Generalist Framework for Panoptic Segmentation of Images and Videos
Ting Chen, Lala Li, Saurabh Saxena, Geoffrey Hinton, David J. Fleet; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9 09-919
Panoptic segmentation assigns semantic and instance ID labels to every pixel of an image. As permutations of instance IDs are also valid solutions, the task req uires learning of high-dimensional one-to-many mapping. As a result, state-of-th e-art approaches use customized architectures and task-specific loss functions. We formulate panoptic segmentation as a discrete data generation problem, withou t relying on inductive bias of the task. A diffusion model is proposed to model panoptic masks, with a simple architecture and generic loss function. By simply adding past predictions as a conditioning signal, our method is capable of model ing video (in a streaming setting) and thereby learns to track object instances automatically. With extensive experiments, we demonstrate that our simple approa ch can perform competitively to state-of-the-art specialist methods in similar s ettings.
****
DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Gener ation Models
Jaemin Cho, Abhay Zala, Mohit Bansal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3043-3054
Recently, DALL-E, a multimodal transformer language model, and its variants incl uding diffusion models have shown high-quality text-to-image generation capabili ties. However, despite the realistic image generation results, there has not bee n a detailed analysis of how to evaluate such models. In this work, we investiga

te the visual reasoning capabilities and social biases of different text-to-imag
e models, covering both multimodal transformer language models and diffusion mod
els. First, we measure three visual reasoning skills: object recognition, object
 counting, and spatial relation understanding. For this, we propose PaintSkills,
 a compositional diagnostic evaluation dataset that measures these skills. Despi
te the high-fidelity image generation capability, a large gap exists between the
 performance of recent models and the upper bound accuracy in object counting an
d spatial relation understanding skills. Second, we assess the gender and skin t
one biases by measuring the gender/skin tone distribution of generated images ac
ross various professions and attributes. We demonstrate that recent text-to-imag
e generation models learn specific biases about gender and skin tone from web im
age-text pairs. We hope our work will help guide future progress in improving te
xt-to-image generation models on visual reasoning skills and learning socially u
nbiased representations.
****
Video OWL-ViT: Temporally-consistent Open-world Localization in Video
Georg Heigold, Matthias Minderer, Alexey Gritsenko, Alex Bewley, Daniel Keysers,
 Mario Lu■i■, Fisher Yu, Thomas Kipf; Proceedings of the IEEE/CVF International
Conference on Computer Vision (ICCV), 2023, pp. 13802-13811
We present an architecture and a training recipe that adapts pretrained open-wor
ld image models to localization in videos. Understanding the open visual world (
without being constrained by fixed label spaces) is crucial for many real-world
vision tasks. Contrastive pre-training on large image-text datasets has recently
 led to significant improvements for image-level tasks. For more structured task
s involving object localization applying pre-trained models is more challenging.
 This is particularly true for video tasks, where task-specific data is limited.
 We show successful transfer of open-world models by building on the OWL-ViT ope
n-vocabulary detection model and adapting it to video by adding a transformer de
coder. The decoder propagates object representations recurrently through time by
 using the output tokens for one frame as the object queries for the next. Our m
odel is end-to-end trainable on video data and enjoys improved temporal consiste
ncy compared to tracking-by-detection baselines, while retaining the open-world
capabilities of the backbone detector. We evaluate our model on the challenging
TAO-OW benchmark and demonstrate that open-world capabilities, learned from larg
e-scale image-text pretraining, can be transferred successfully to open-world lo
calization across diverse videos.
****
Few Shot Font Generation Via Transferring Similarity Guided Global Style and Qua
ntization Local Style
Wei Pan, Anna Zhu, Xinyu Zhou, Brian Kenji Iwana, Shilin Li; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19506-195
16
Automatic few-shot font generation (AFFG), aiming at generating new fonts with o
nly a few glyph references, reduces the labor cost of manually designing fonts.
However, the traditional AFFG paradigm of style-content disentanglement cannot c
apture the diverse local details of different fonts. So, many component-based ap
proaches are proposed to tackle this problem. The issue with component-based app
roaches is that they usually require special pre-defined glyph components, e.g.,
 strokes and radicals, which is infeasible for AFFG of different languages. In t
his paper, we present a novel font generation approach by aggregating styles fro
m character similarity-guided global features and stylized component-level repre
sentations. We calculate the similarity scores of the target character and the r
eferenced samples by measuring the distance along the corresponding channels fro
m the content features, and assigning them as the weights for aggregating the gl
obal style features. To better capture the local styles, a cross-attention-based
 style transfer module is adopted to transfer the styles of reference glyphs to
the components, where the components are self-learned discrete latent codes thro
ugh vector quantization without manual definition. With these designs, our AFFG
method could obtain a complete set of component-level style representations, and
 also control the global glyph characteristics. The experimental results reflect

the effectiveness and generalization of the proposed method on different linguistic scripts, and also show its superiority when compared with other state-of-the-art methods. The source code can be found at https://github.com/awei669/VQ-Font.
****
Differentiable Transportation Pruning
Yunqiang Li, Jan C. van Gemert, Torsten Hoefler, Bert Moons, Evangelos Eleftheriou, Bram-Ernst Verhoef; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16957-16967
Deep learning algorithms are increasingly employed at the edge. However, edge devices are resource constrained and thus require efficient deployment of deep neural networks. Pruning methods are a key tool for edge deployment as they can improve storage, compute, memory bandwidth, and energy usage. In this paper we propose a novel accurate pruning technique that allows precise control over the output network size. Our method uses an efficient optimal transportation scheme which we make end-to-end differentiable and which automatically tunes the exploration-exploitation behavior of the algorithm to find accurate sparse sub-networks. We show that our method achieves state-of-the-art performance compared to previous pruning methods on 3 different datasets, using 5 different models, across a wide range of pruning ratios, and with two types of sparsity budgets and pruning granularities.
****
Physics-Driven Turbulence Image Restoration with Stochastic Refinement
Ajay Jaiswal, Xingguang Zhang, Stanley H. Chan, Zhangyang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12170-12181
Image distortion by atmospheric turbulence is a stochastic degradation, which is a critical problem in long-range optical imaging systems. A number of research has been conducted during the past decades, including model-based and emerging deep-learning solutions with the help of synthetic data. Although fast and physics-grounded simulation tools have been introduced to help the deep-learning models adapt to real-world turbulence conditions recently, the training of such models only relies on the synthetic data and ground truth pairs. This paper proposes the Physics-integrated Restoration Network (PiRN) to bring the physics-based simulator directly into the training process to help the network to disentangle the stochasticity from the degradation and the underlying image. Furthermore, to overcome the "average effect" introduced by deterministic models and the domain gap between the synthetic and real-world degradation, we further introduce PiRN with Stochastic Refinement (PiRN-SR) to boost its perceptual quality. Overall, our PiRN and PiRN-SR improve the generalization to real-world unknown turbulence conditions and provide a state-of-the-art restoration in both pixel-wise accuracy and perceptual quality.
****
Enhancing Non-line-of-sight Imaging via Learnable Inverse Kernel and Attention Mechanisms
Yanhua Yu, Siyuan Shen, Zi Wang, Binbin Huang, Yuehan Wang, Xingyue Peng, Suan Xia, Ping Liu, Ruiqian Li, Shiying Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10563-10573
Recovering information from non-line-of-sight (NLOS) imaging is a computationally-intensive inverse problem. Most physics-based NLOS imaging methods address the complexity of this problem by assuming three-bounce reflections and no self-occlusion. However, these assumptions may break down for objects with large depth variations, preventing physics-based algorithms from accurately reconstructing the details and high-frequency information. On the other hand, while learning-based methods can avoid these assumptions, they may struggle to reconstruct details without specific designs due to the spectral bias of neural networks. To overcome these issues, we propose a novel approach that enhances physics-based NLOS imaging methods by introducing a learnable inverse kernel in the Fourier domain and using an attention mechanism to improve the neural network to learn high-frequency information. Our method is evaluated on publicly available and new synthetic

datasets, demonstrating its commendable performance compared to prior physics-based and learning-based methods, especially for objects with large depth variations. Moreover, our approach generalizes well to real data and can be applied to tasks such as classification and depth reconstruction. We will make our code and dataset publicly available: https://sci2020.github.io.
****
DECO: Dense Estimation of 3D Human-Scene Contact In The Wild
Shashank Tripathi, Agniv Chatterjee, Jean-Claude Passy, Hongwei Yi, Dimitrios Tzionas, Michael J. Black; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8001-8013
Understanding how humans use physical contact to interact with the world is key to enabling human-centric artificial intelligence. While inferring 3D contact is crucial for modeling realistic and physically-plausible human-object interactions, existing methods either focus on 2D, consider body joints rather than the surface, use coarse 3D body regions, or do not generalize to in-the-wild images. In contrast, we focus on inferring dense, 3D contact between the full body surface and objects in arbitrary images. To achieve this, we first collect DAMON, a new dataset containing dense vertex-level contact annotations paired with RGB images containing complex human-object and human-scene contact. Second, we train DECO, a novel 3D contact detector that uses both body-part-driven and scene-context-driven attention to estimate vertex-level contact on the SMPL body. DECO builds on the insight that human observers recognize contact by reasoning about the contacting body parts, their proximity to scene objects, and the surrounding scene context.
We perform extensive evaluations of our detector on DAMON as well as on the RICH and BEHAVE datasets. We significantly outperform existing SOTA methods across all benchmarks.
We also show qualitatively that DECO generalizes well to diverse and challenging real-world human interactions in natural images. The code, data, and models are available at https://deco.is.tue.mpg.de.
****
Scale-Aware Modulation Meet Transformer
Weifeng Lin, Ziheng Wu, Jiayu Chen, Jun Huang, Lianwen Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6015-6026
This paper presents a new vision Transformer, ScaleAware Modulation Transformer (SMT), that can handle various downstream tasks efficiently by combining the convolutional network and vision Transformer. The proposed ScaleAware Modulation (SAM) in the SMT includes two primary novel designs. Firstly, we introduce the Multi-Head Mixed Convolution (MHMC) module, which can capture multiscale features and expand the receptive field. Secondly, we propose the Scale-Aware Aggregation (SAA) module, which is lightweight but effective, enabling information fusion across different heads. By leveraging these two modules, convolutional modulation is further enhanced. Furthermore, in contrast to prior works that utilized modulations throughout all stages to build an attention-free network, we propose an Evolutionary Hybrid Network (EHN), which can effectively simulate the shift from capturing local to global dependencies as the network becomes deeper, resulting in superior performance. Extensive experiments demonstrate that SMT significantly outperforms existing state-of-the-art models across a wide range of visual tasks. Specifically, SMT with 11.5M / 2.4GFLOPs and 32M / 7.7GFLOPs can achieve 82.2% and 84.3% top-1 accuracy on ImageNet-1K, respectively. After pretrained on ImageNet-22K in 224x224 resolution, it attains 87.1% and 88.1% top-1 accuracy when finetuned with resolution 224x224 and 384x384 , respectively. For object detection with Mask R-CNN, the SMT base trained with 1x and 3x schedule outperforms the Swin Transformer counterpart by 4.2 and 1.3 mAP on COCO, respectively. For semantic segmentation with UPerNet, the SMT base test at single- and multi-scale surpasses Swin by 2.0 and 1.1 mIoU respectively on the ADE20K. Our code is available at https://github.com/AFeng-x/SMT.
****
Large Selective Kernel Network for Remote Sensing Object Detection
Yuxuan Li, Qibin Hou, Zhaohui Zheng, Ming-Ming Cheng, Jian Yang, Xiang Li; Proce

edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16794-16805
Recent research on remote sensing object detection has largely focused on improving the representation of oriented bounding boxes but has overlooked the unique prior knowledge presented in remote sensing scenarios. Such prior knowledge can be useful because tiny remote sensing objects may be mistakenly detected without referencing a sufficiently long-range context, which can vary for different objects. This paper considers these priors and proposes the lightweight Large Selective Kernel Network (LSKNet). LSKNet can dynamically adjust its large spatial receptive field to better model the ranging context of various objects in remote sensing scenarios. To our knowledge, large and selective kernel mechanisms have not been previously explored in remote sensing object detection. Without bells and whistles, our lightweight LSKNet sets new state-of-the-art scores on standard benchmarks, i.e., HRSC2016 (98.46% mAP), DOTA-v1.0 (81.85% mAP), and FAIR1M-v1.0 (47.87% mAP).
****
PlaneRecTR: Unified Query Learning for 3D Plane Recovery from a Single View
Jingjia Shi, Shuaifeng Zhi, Kai Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9377-9386
3D plane recovery from a single image can usually be divided into several subtasks of plane detection, segmentation, parameter estimation and possibly depth estimation. Previous works tend to solve it by either extending the RCNN-based segmentation network or the dense pixel embedding-based clustering framework. However, none of them tried to integrate above related subtasks into a unified framework but treated them separately and sequentially, which we suspect is potentially a main source of performance limitation for existing approaches. Motivated by this finding and the success of query-based learning in enriching reasoning among semantic entities, in this paper, we propose PlaneRecTR, a Transformer-based architecture, which for the first time unifies all subtasks related to single-view plane recovery with a single compact model. Extensive quantitative and qualitative experiments demonstrate that our proposed unified learning achieves mutual benefits across subtasks, obtaining a new state-of-the-art performance on public ScanNet and NYUv2-Plane datasets.
****
EigenTrajectory: Low-Rank Descriptors for Multi-Modal Trajectory Forecasting
Inhwan Bae, Jean Oh, Hae-Gon Jeon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10017-10029
Capturing high-dimensional social interactions and feasible futures is essential for predicting trajectories. To address this complex nature, several attempts have been devoted to reducing the dimensionality of the output variables via parametric curve fitting such as the Bezier curve and B-spline function. However, these functions, which originate in computer graphics fields, are not suitable to account for socially acceptable human dynamics. In this paper, we present EigenTrajectory (ET), a trajectory prediction approach that uses a novel trajectory descriptor to form a compact space, known here as ET space, in place of Euclidean space, for representing pedestrian movements. We first reduce the complexity of the trajectory descriptor via a low-rank approximation. We transform the pedestrians' history paths into our ET space represented by spatio-temporal principle components, and feed them into off-the-shelf trajectory forecasting models. The inputs and outputs of the models as well as social interactions are all gathered and aggregated in the corresponding ET space. Lastly, we propose a trajectory anchor-based refinement method to cover all possible futures in the proposed ET. Extensive experiments demonstrate that our EigenTrajectory predictor can significantly improve both the prediction accuracy and reliability of existing trajectory forecasting models on public benchmarks, indicating that the proposed descriptor is suited to represent pedestrian behaviors. Code is publicly available at https://github.com/inhwanbae/EigenTrajectory.
****
I-ViT: Integer-only Quantization for Efficient Vision Transformer Inference
Zhikai Li, Qingyi Gu; Proceedings of the IEEE/CVF International Conference on Co

mputer Vision (ICCV), 2023, pp. 17065-17075
Vision Transformers (ViTs) have achieved state-of-the-art performance on various
 computer vision applications. However, these models have considerable storage a
nd computational overheads, making their deployment and efficient inference on e
dge devices challenging. Quantization is a promising approach to reducing model
complexity, and the dyadic arithmetic pipeline can allow the quantized models to
 perform efficient integer-only inference. Unfortunately, dyadic arithmetic is b
ased on the homogeneity condition in convolutional neural networks, which is not
 applicable to the non-linear components in ViTs, making integer-only inference
of ViTs an open issue. In this paper, we propose I-ViT, an integer-only quantiza
tion scheme for ViTs, to enable ViTs to perform the entire computational graph o
f inference with integer arithmetic and bit-shifting, and without any floating-p
oint arithmetic. In I-ViT, linear operations (e.g., MatMul and Dense) follow the
 integer-only pipeline with dyadic arithmetic, and non-linear operations (e.g.,
Softmax, GELU, and LayerNorm) are approximated by the proposed light-weight inte
ger-only arithmetic methods. More specifically, I-ViT applies the proposed Shift
max and ShiftGELU, which are designed to use integer bit-shifting to approximate
 the corresponding floating-point operations. We evaluate I-ViT on various bench
mark models and the results show that integer-only INT8 quantization achieves co
mparable (or even slightly higher) accuracy to the full-precision (FP) baseline.
 Furthermore, we utilize TVM for practical hardware deployment on the GPU's inte
ger arithmetic units, achieving 3.72 4.11x inference speedup compared to the FP
model. Code of both Pytorch and TVM is released at
 https://github.com/zkkli/I-ViT.
****
SUMMIT: Source-Free Adaptation of Uni-Modal Models to Multi-Modal Targets
Cody Simons, Dripta S. Raychaudhuri, Sk Miraj Ahmed, Suya You, Konstantinos Kary
dis, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF International Conference
 on Computer Vision (ICCV), 2023, pp. 1239-1249
Scene understanding using multi-modal data is necessary in many applications, e.
g., autonomous navigation. To achieve this in a variety of situations, existing
models must be able to adapt to shifting data distributions without arduous data
 annotation. Current approaches assume that the source data is available during
adaptation and that the source consists of paired multi-modal data. Both these a
ssumptions may be problematic for many applications. Source data may not be avai
lable due to privacy, security, or economic concerns. Assuming the existence of
paired multi-modal data for training also entails significant data collection co
sts and fails to take advantage of widely available freely distributed pre-train
ed uni-modal models. In this work, we relax both of these assumptions by address
ing the problem of adapting a set of models trained independently on uni-modal d
ata to a target domain consisting of unlabeled multi-modal data, without having
access to the original source dataset. Our proposed approach solves this problem
 through a switching framework which automatically chooses between two complemen
tary methods of cross-modal pseudo-label fusion -- agreement filtering and entro
py weighting -- based on the estimated domain gap. We demonstrate our work on th
e semantic segmentation problem. Experiments across seven challenging adaptation
 scenarios verify the efficacy of our approach, achieving results comparable to,
 and in some cases outperforming, methods which assume access to source data. Ou
r method achieves an improvement in mIoU of up to 12% over competing baselines.
Our code is publicly available at https://github.com/csimo005/SUMMIT.
****
Learning a More Continuous Zero Level Set in Unsigned Distance Fields through Le
vel Set Projection
Junsheng Zhou, Baorui Ma, Shujuan Li, Yu-Shen Liu, Zhizhong Han; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3181-
3192
Latest methods represent shapes with open surfaces using unsigned distance funct
ions (UDFs). They train neural networks to learn UDFs and reconstruct surfaces w
ith the gradients around the zero level set of the UDF. However, the differentia
l networks struggle from learning the zero level set where the UDF is not differ

entiable, which leads to large errors on unsigned distances and gradients around the zero level set, resulting in highly fragmented and discontinuous surfaces. To resolve this problem, we propose to learn a more continuous zero level set in UDFs with level set projections. Our insight is to guide the learning of zero level set using the rest non-zero level sets via a projection procedure. Our idea is inspired from the observations that the non-zero level sets are much smoother and more continuous than the zero level set. We pull the non-zero level sets onto the zero level set with gradient constraints which align gradients over different level sets and correct unsigned distance errors on the zero level set, leading to a smoother and more continuous unsigned distance field. We conduct comprehensive experiments in surface reconstruction for point clouds, real scans or depth maps, and further explore the performance in unsupervised point cloud upsampling and unsupervised point normal estimation with the learned UDF, which demonstrate our non-trivial improvements over the state-of-the-art methods. Code is available at https://github.com/junshengzhou/LevelSetUDF.
****
Video-FocalNets: Spatio-Temporal Focal Modulation for Video Action Recognition
Syed Talal Wasim, Muhammad Uzair Khattak, Muzammal Naseer, Salman Khan, Mubarak Shah, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13778-13789
Recent video recognition models utilize Transformer models for long-range spatio-temporal context modeling. Video transformer designs are based on self-attention that can model global context at a high computational cost. In comparison, convolutional designs for videos offer an efficient alternative but lack long-range dependency modeling. Towards achieving the best of both designs, this work proposes Video-FocalNet, an effective and efficient architecture for video recognition that models both local and global contexts. Video-FocalNet is based on a spatio-temporal focal modulation architecture that reverses the interaction and aggregation steps of self-attention for better efficiency. Further, the aggregation step and the interaction step are both implemented using efficient convolution and element-wise multiplication operations that are computationally less expensive than their self-attention counterparts on video representations. We extensively explore the design space of focal modulation-based spatio-temporal context modeling and demonstrate our parallel spatial and temporal encoding design to be the optimal choice. Video-FocalNets perform favorably well against the state-of-the-art transformer-based models for video recognition on five large-scale datasets (Kinetics-400, Kinetics-600, SS-v2, Diving-48, and ActivityNet-1.3) at a lower computational cost. Our code/models are released at https://github.com/TalalWasim/Video-FocalNets.
****
To Adapt or Not to Adapt? Real-Time Adaptation for Semantic Segmentation
Marc Botet Colomer, Pier Luigi Dovesi, Theodoros Panagiotakopoulos, Joao Frederico Carvalho, Linus Härenstam-Nielsen, Hossein Azizpour, Hedvig Kjellström, Daniel Cremers, Matteo Poggi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16548-16559
The goal of Online Domain Adaptation for semantic segmentation is to handle unforeseeable domain changes that occur during deployment, like sudden weather events. However, the high computational costs associated with brute-force adaptation make this paradigm unfeasible for real-world applications. In this paper we propose HAMLET, a Hardware-Aware Modular Least Expensive Training framework for real-time domain adaptation. Our approach includes a hardware-aware back-propagation orchestration agent (HAMT) and a dedicated domain-shift detector that enables active control over when and how the model is adapted (LT). Thanks to these advancements, our approach is capable of performing semantic segmentation while simultaneously adapting at more than 29FPS on a single consumer-grade GPU. Our framework's encouraging accuracy and speed trade-off is demonstrated on OnDA and SHIFT benchmarks through experimental results.
****
Hidden Biases of End-to-End Driving Models
Bernhard Jaeger, Kashyap Chitta, Andreas Geiger; Proceedings of the IEEE/CVF Int

ernational Conference on Computer Vision (ICCV), 2023, pp. 8240-8249
End-to-end driving systems have recently made rapid progress, in particular on C
ARLA. Independent of their major contribution, they introduce changes to minor s
ystem components. Consequently, the source of improvements is unclear. We identi
fy two biases that recur in nearly all state-of-the-art methods and are critical
 for the observed progress on CARLA: (1) lateral recovery via a strong inductive
 bias towards target point following, and (2) longitudinal averaging of multimod
al waypoint predictions for slowing down. We investigate the drawbacks of these
biases and identify principled alternatives. By incorporating our insights, we d
evelop TF++, a simple end-to-end method that ranks first on the Longest6 and LAV
 benchmarks, gaining 11 driving score over the best prior work on Longest6.
****
HairNeRF: Geometry-Aware Image Synthesis for Hairstyle Transfer
Seunggyu Chang, Gihoon Kim, Hayeon Kim; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2023, pp. 2448-2458
We propose a novel hairstyle transferred image synthesis method considering the
underlying head geometry of two input images. In traditional GAN-based methods,
transferring hairstyle from one image to the other often makes the synthesized r
esult awkward due to differences in pose, shape, and size of heads. To resolve t
his, we utilize neural rendering by registering two input heads in the volumetri
c space to make a transferred hairstyle fit on the head of a target image. Becau
se of the geometric nature of neural rendering, our method can render view varyi
ng images of synthesized results from a single transfer process without causing
distortion from which extant hairstyle transfer methods built upon traditional G
AN-based generators suffer. We verify that our method surpasses other baselines
in view of preserving the identity and hairstyle of two input images when synthe
sizing a hairstyle transferred image rendered at any point of view.
****
Strivec: Sparse Tri-Vector Radiance Fields
Quankai Gao, Qiangeng Xu, Hao Su, Ulrich Neumann, Zexiang Xu; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17569-17
579
We propose Strivec, a novel neural representation that models a 3D scene as a ra
diance field with sparsely distributed and compactly factorized local tensor fea
ture grids. Our approach leverages tensor decomposition, following the recent wo
rk TensoRF, to model the tensor grids. In contrast to TensoRF which uses a globa
l tensor and focuses on their vector-matrix decomposition, we propose to utilize
 a cloud of local tensors and apply the classic CANDECOMP/PARAFAC (CP) decomposi
tion to factorize each tensor into triple vectors that express local feature dis
tributions along spatial axes and compactly encode a local neural field. We also
 apply multi-scale tensor grids to discover the geometry and appearance commonal
ities and exploit spatial coherence with the tri-vector factorization at multipl
e local scales. The final radiance field properties are regressed by aggregating
 neural features from multiple local tensors across all scales. Our tri-vector t
ensors are sparsely distributed around the actual scene surface, discovered by a
 fast coarse reconstruction, leveraging the sparsity of a 3D scene. We demonstra
te that our model can achieve better rendering quality while using significantly
 fewer parameters than previous methods, including TensoRF and Instant-NGP.
****
Multiscale Representation for Real-Time Anti-Aliasing Neural Rendering
Dongting Hu, Zhenkai Zhang, Tingbo Hou, Tongliang Liu, Huan Fu, Mingming Gong; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 17772-17783
The rendering scheme in neural radiance field (NeRF) is effective in rendering a
 pixel by casting a ray into the scene. However, NeRF yields blurred rendering r
esults when the training images are captured at non-uniform scales, and produces
 aliasing artifacts if the test images are taken in distant views. To address th
is issue, Mip-NeRF proposes a multiscale representation as a conical frustum to
encode scale information. Nevertheless, this approach is only suitable for offli
ne rendering since it relies on integrated positional encoding (IPE) to query a

multilayer perceptron (MLP). To overcome this limitation, we propose mip voxel grids (Mip-VoG), an explicit multiscale representation with a deferred architecture for real-time anti-aliasing rendering. Our approach includes a density Mip-VoG for scene geometry and a feature Mip-VoG with a small MLP for view-dependent color. Mip-VoG represents scene scale using the level of detail (LOD) derived from ray differentials and uses quadrilinear interpolation to map a queried 3D location to its features and density from two neighboring down-sampled voxel grids. To our knowledge, our approach is the first to offer multiscale training and real-time anti-aliasing rendering simultaneously. We conducted experiments on multiscale dataset, results show that our approach outperforms state-of-the-art real-time rendering baselines.
****
Borrowing Knowledge From Pre-trained Language Model: A New Data-efficient Visual Learning Paradigm
Wenxuan Ma, Shuang Li, JinMing Zhang, Chi Harold Liu, Jingxuan Kang, Yulin Wang, Gao Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18786-18797
The development of vision models for real-world applications is hindered by the challenge of annotated data scarcity, which has necessitated the adoption of data-efficient visual learning techniques such as semi-supervised learning. Unfortunately, the prevalent cross-entropy supervision is limited by its focus on category discrimination while disregarding the semantic connection between concepts, which ultimately results in the suboptimal exploitation of scarce labeled data. To address this issue, this paper presents a novel approach that seeks to leverage linguistic knowledge for data-efficient visual learning. The proposed approac, BorLan, Borrows knowledge from off-the-shelf pretrained Language models that are already endowed with rich semantics extracted from large corpora, to compensate the semantic deficiency due to limited annotation in visual training. Specifically, we design a distribution alignment objective, which guides the vision model to learn both semantic-aware and domain-agnostic representations for the task through linguistic knowledge. One significant advantage of this paradigm is its flexibility in combining various visual and linguistic models. Extensive experiments on semi-supervised learning, single domain generalization and few-shot learning validate its effectiveness.
****
GETAvatar: Generative Textured Meshes for Animatable Human Avatars
Xuanmeng Zhang, Jianfeng Zhang, Rohan Chacko, Hongyi Xu, Guoxian Song, Yi Yang, Jiashi Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2273-2282
We study the problem of 3D-aware full-body human generation, aiming at creating animatable human avatars with high-quality textures and geometries. Generally, two challenges remain in this field: i) existing methods struggle to generate geometries with rich realistic details such as the wrinkles of garments; ii) they typically utilize volumetric radiance fields and neural renderers in the synthesis process, making high-resolution rendering non-trivial. To overcome these problems, we propose GETAvatar, a Generative model that directly generates Explicit Textured 3D meshes
for animatable human Avatar, with photo-realistic appearance and fine geometric details. Specifically, we first design an articulated 3D human representation with explicit surface modeling, and enrich the generated humans with realistic surface details by learning from the 2D normal maps of 3D scan data. Second, with the explicit mesh representation, we can use a rasterization-based renderer to perform surface rendering, allowing us to achieve high-resolution image generation efficiently. Extensive experiments demonstrate that GETAvatar achieves state-of-the-art performance on 3D-aware human generation both in appearance and geometry quality. Notably, GETAvatar cangenerate images at 512x512 resolution with 17FPS and 1024x1024 resolution with 14FPS, improving upon previous methods by 2x. Our code and models will be available.
****
Tracking without Label: Unsupervised Multiple Object Tracking via Contrastive Si

milarity Learning
Sha Meng, Dian Shao, Jiacheng Guo, Shan Gao; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2023, pp. 16264-16273
Unsupervised learning is a challenging task due to the lack of labels. Multiple
Object Tracking (MOT), which inevitably suffers from mutual object interference,
 occlusion, etc., is even more difficult without label supervision. In this pape
r, we explore the latent consistency of sample features across video frames and
propose an Unsupervised Contrastive Similarity Learning method, named UCSL, incl
uding three contrast modules: self-contrast, cross-contrast, and ambiguity contr
ast. Specifically, i) self-contrast uses intra-frame direct and inter-frame indi
rect contrast to obtain discriminative representations by maximizing self-simila
rity. ii) Cross-contrast aligns cross- and continuous-frame matching results, mi
tigating the persistent negative effect caused by object occlusion. And iii) amb
iguity contrast matches ambiguous objects with each other to further increase th
e certainty of subsequent object association through an implicit manner. On exis
ting benchmarks, our method outperforms the existing unsupervised methods using
only limited help from ReID head, and even provides higher accuracy than lots of
 fully supervised methods.
****
PIDRo: Parallel Isomeric Attention with Dynamic Routing for Text-Video Retrieval
Peiyan Guan, Renjing Pei, Bin Shao, Jianzhuang Liu, Weimian Li, Jiaxi Gu, Hang X
u, Songcen Xu, Youliang Yan, Edmund Y. Lam; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 11164-11173
Text-video retrieval is a fundamental task with high practical value in multi-mo
dal research. Inspired by the great success of pre-trained image-text models wit
h large-scale data, such as CLIP, many methods are proposed to transfer the stro
ng representation learning capability of CLIP to text-video retrieval. However,
due to the modality difference between videos and images, how to effectively ada
pt CLIP to the video domain is still underexplored. In this paper, we investigat
e this problem from two aspects. First, we enhance the transferred image encoder
 of CLIP for fine-grained video understanding in a seamless fashion. Second, we
conduct fine-grained contrast between videos and texts from both model improveme
nt and loss design. Particularly, we propose a fine-grained contrastive model eq
uipped with parallel isomeric attention and dynamic routing, namely PIDRo, for t
ext-video retrieval. The parallel isomeric attention module is used as the video
 encoder, which consists of two parallel branches modeling the spatial-temporal
information of videos from both patch and frame levels. The dynamic routing modu
le is constructed to enhance the text encoder of CLIP, generating informative wo
rd representations by distributing the fine-grained information to the related w
ord tokens within a sentence. Such model design provides us with informative pat
ch, frame and word representations. We then conduct token-wise interaction upon
them. With the enhanced encoders and the token-wise loss, we are able to achieve
 finer-grained text-video alignment and more accurate retrieval. PIDRo obtains s
tate-of-the-art performance over various text-video retrieval benchmarks, includ
ing MSR-VTT, MSVD, LSMDC, DiDeMo and ActivityNet.
****
Re-mine, Learn and Reason: Exploring the Cross-modal Semantic Correlations for L
anguage-guided HOI detection
Yichao Cao, Qingfei Tang, Feng Yang, Xiu Su, Shan You, Xiaobo Lu, Chang Xu; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023
, pp. 23492-23503
Human-Object Interaction (HOI) detection is a challenging computer vision task t
hat requires visual models to address the complex interactive relationship betwe
en humans and objects and predict <human, action, object> triplets. Despite the
challenges posed by the numerous interaction combinations, they also offer oppor
tunities for multi-modal learning of visual texts. In this paper, we present a s
ystematic and unified framework (RmLR) that enhances HOI detection by incorporat
ing structured text knowledge. Firstly, we qualitatively and quantitatively anal
yze the loss of interaction information in the two-stage HOI detector and propos
e a re-mining strategy to generate more comprehensive visual representation. Sec

ondly, we design more fine-grained sentence- and word-level alignment and knowle dge transfer strategies to effectively address the many-to-many matching problem between multiple interactions and multiple texts. These strategies alleviate th e matching confusion problem that arises when multiple interactions occur simult aneously, thereby improving the effectiveness of the alignment process. Finally, HOI reasoning by visual features augmented with textual knowledge substantially improves the understanding of interactions. Experimental results illustrate the effectiveness of our approach, where state-of-the-art performance is achieved o n public benchmarks.
****
Strata-NeRF : Neural Radiance Fields for Stratified Scenes
Ankit Dhiman, R Srinath, Harsh Rangwani, Rishubh Parihar, Lokesh R Boregowda, Sr inath Sridhar, R Venkatesh Babu; Proceedings of the IEEE/CVF International Confe rence on Computer Vision (ICCV), 2023, pp. 17603-17614
Neural Radiance Fields (NeRF) approaches learn the underlying 3D representation of a scene and generate photo-realistic novel views with high fidelity. However, most proposed settings concentrate on 3D modelling a single object or a single level of a scene. However, in the real world, a person captures a structure at m ultiple levels, resulting in layered capture. For example, tourists usually capt ure a monument's exterior structure before capturing the inner structure. Modell ing such scenes in 3D with seamless switching between levels can drastically imp rove the Virtual Reality (VR) experience. However, most of the existing techniqu es struggle in modelling such scenes. Hence, we propose Strata-NeRF, a single ra diance field that can implicitly learn the 3D representation of outer, inner, an d subsequent levels. Strata-NeRF achieves this by conditioning the NeRFs on Vect or Quantized (VQ) latents which allows sudden changes in scene structure with ch anges in levels due to their discrete nature. We first investigate the proposed approach's effectiveness by modelling a novel multilayered synthetic dataset com prising diverse scenes and then further validate its generalization on the real- world RealEstate dataset. We find that Strata-NeRF effectively models the scene structure, minimizes artefacts and synthesizes high-fidelity views compared to e xisting state-of-the-art approaches in the literature.
****
StylerDALLE: Language-Guided Style Transfer Using a Vector-Quantized Tokenizer o f a Large-Scale Generative Model
Zipeng Xu, Enver Sangineto, Nicu Sebe; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7601-7611
Despite the progress made in the style transfer task, most previous work focus o n transferring only relatively simple features like color or texture, while miss ing more abstract concepts such as overall art expression or painter-specific tr aits. However, these abstract semantics can be captured by models like DALL-E or CLIP, which have been trained using huge datasets of images and textual documen ts. In this paper, we propose StylerDALLE, a style transfer method that exploits both of these models and uses natural language to describe abstract art styles. Specifically, we formulate the language-guided style transfer task as a non-aut oregressive token sequence translation, i.e., from input content image to output stylized image, in the discrete latent space of a large-scale pretrained vector -quantized tokenizer, e.g., the discrete variational auto-encoder (dVAE) of DALL -E. To incorporate style information, we propose a Reinforcement Learning strate gy with CLIP-based language supervision that ensures stylization and content pre servation simultaneously. Experimental results demonstrate the superiority of ou r method, which can effectively transfer art styles using language instructions at different granularities. Code is available at https://github.com/zipengxuc/St ylerDALLE.
****
3D-aware Blending with Generative NeRFs
Hyunsu Kim, Gayoung Lee, Yunjey Choi, Jin-Hwa Kim, Jun-Yan Zhu; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22906- 22918
Image blending aims to combine multiple images seamlessly. It remains challengin

g for existing 2D-based methods, especially when input images are misaligned due to differences in 3D camera poses and object shapes. To tackle these issues, we propose a 3D-aware blending method using generative Neural Radiance Fields (NeRF), including two key components: 3D-aware alignment and 3D-aware blending. For 3D-aware alignment, we first estimate the camera pose of the reference image with respect to generative NeRFs and then perform pose alignment for objects. To further leverage 3D information of the generative NeRF, we propose 3D-aware blending that utilizes volume density and blends on the NeRF's latent space, rather than raw pixel space. Collectively, our method outperforms existing 2D baselines, as validated by extensive quantitative and qualitative evaluations with FFHQ and AFHQ-Cat.
****
Multi-Modal Gated Mixture of Local-to-Global Experts for Dynamic Image Fusion
Bing Cao, Yiming Sun, Pengfei Zhu, Qinghua Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23555-23564
Infrared and visible image fusion aims to integrate comprehensive information from multiple sources to achieve superior performances on various practical tasks, such as detection, over that of a single modality. However, most existing methods directly combined the texture details and object contrast of different modalities, ignoring the dynamic changes in reality, which diminishes the visible texture in good lighting conditions and the infrared contrast in low lighting conditions. To fill this gap, we propose a dynamic image fusion framework with a multi-modal gated mixture of local-to-global experts, termed MoE-Fusion, to dynamically extract effective and comprehensive information from the respective modalities. Our model consists of a Mixture of Local Experts (MoLE) and a Mixture of Global Experts (MoGE) guided by a multi-modal gate. The MoLE performs specialized learning of multi-modal local features, prompting the fused images to retain the local information in a sample-adaptive manner, while the MoGE focuses on the global information that complements the fused image with overall texture detail and contrast. Extensive experiments show that our MoE-Fusion outperforms state-of-the-art methods in preserving multi-modal image texture and contrast through the local-to-global dynamic learning paradigm, and also achieves superior performance on detection tasks. Our code is available: https://github.com/SunYM2020/MoE-Fusion.
****
Deep Image Harmonization with Learnable Augmentation
Li Niu, Junyan Cao, Wenyan Cong, Liqing Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7482-7491
The goal of image harmonization is adjusting the foreground appearance in a composite image to make the whole image harmonious. To construct paired training images, existing datasets adopt different ways to adjust the illumination statistics of foregrounds of real images to produce synthetic composite images. However, different datasets have considerable domain gap and the performances on small-scale datasets are limited by insufficient training data. In this work, we explore learnable augmentation to enrich the illumination diversity of small-scale datasets for better harmonization performance. In particular, our designed SYthetic COmposite Network (SycoNet) takes in a real image with foreground mask and a random vector to learn suitable color transformation, which is applied to the foreground of this real image to produce a synthetic composite image. Comprehensive experiments demonstrate the effectiveness of our proposed learnable augmentation for image harmonization. The code of SycoNet is released at https://github.com/bcmi/SycoNet-Adaptive-Image-Harmonization.
****
DELFlow: Dense Efficient Learning of Scene Flow for Large-Scale Point Clouds
Chensheng Peng, Guangming Wang, Xian Wan Lo, Xinrui Wu, Chenfeng Xu, Masayoshi Tomizuka, Wei Zhan, Hesheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16901-16910
Point clouds are naturally sparse, while image pixels are dense. The inconsistency limits feature fusion from both modalities for point-wise scene flow estimation. Previous methods rarely predict scene flow from the entire point clouds of t

he scene with one-time inference due to the memory inefficiency and heavy overhe
ad from distance calculation and sorting involved in commonly used farthest poin
t sampling, KNN, and ball query algorithms for local feature aggregation. To mit
igate these issues in scene flow learning, we regularize raw points to a dense f
ormat by storing 3D coordinates in 2D grids. Unlike the sampling operation commo
nly used in existing works, the dense 2D representation 1) preserves most points
 in the given scene, 2) brings in a significant boost of efficiency, and 3) elim
inates the density gap between points and pixels, allowing us to perform effecti
ve feature fusion. We also present a novel warping projection technique to allev
iate the information loss problem resulting from the fact that multiple points c
ould be mapped into one grid during projection when computing cost volume. Suffi
cient experiments demonstrate the efficiency and effectiveness of our method, ou
tperforming the prior-arts on the FlyingThings3D and KITTI dataset.
****
RFD-ECNet: Extreme Underwater Image Compression with Reference to Feature Dictio
nary
Mengyao Li, Liquan Shen, Peng Ye, Guorui Feng, Zheyin Wang; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12980-1298
9
Thriving underwater applications demand efficient extreme compression technology
 to realize the transmission of underwater images (UWIs) in very narrow underwat
er bandwidth. However, existing image compression methods achieve inferior perfo
rmance on UWIs because they do not consider the characteristics of UWIs: (1) Mul
tifarious underwater styles of color shift and distance-dependent clarity, cause
d by the unique underwater physical imaging; (2) Massive redundancy between diff
erent UWIs, caused by the fact that different UWIs contain several common ocean
objects, which have plenty of similarities in structures and semantics. To remov
e redundancy among UWIs, we first construct an exhaustive underwater multi-scale
 feature dictionary to provide coarse-to-fine reference features for UWI compres
sion. Subsequently, an extreme UWI compression network with reference to the fea
ture dictionary (RFD-ECNet) is creatively proposed, which utilizes feature match
 and reference feature variant to significantly remove redundancy among UWIs. To
 align the multifarious underwater styles and improve the accuracy of feature ma
tch, an underwater style normalized block (USNB) is proposed, which utilizes und
erwater physical priors extracted from the underwater physical imaging model to
normalize the underwater styles of dictionary features toward the input. Moreove
r, a reference feature variant module (RFVM) is designed to adaptively morph the
 reference features, improving the similarity between the reference and input fe
atures. Experimental results on four UWI datasets show that our RFD-ECNet is the
 first work that achieves a significant BD-rate saving of 31% over the most adva
nced VVC.
****
E^2VPT: An Effective and Efficient Approach for Visual Prompt Tuning
Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, Dongfang
 Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (I
CCV), 2023, pp. 17491-17502
As the size of transformer-based models continues to grow, fine-tuning these lar
ge-scale pre-trained vision models for new tasks has become increasingly paramet
er-intensive. Parameter-efficient learning has been developed to reduce the numb
er of tunable parameters during fine-tuning. Although these methods show promisi
ng results, there is still a significant performance gap compared to full fine-t
uning. To address this challenge, we propose an Effective and Efficient Visual P
rompt Tuning (E^2VPT) approach for large-scale transformer-based model adaptatio
n. Specifically, we introduce a set of learnable key-value prompts and visual pr
ompts into self-attention and input layers, respectively, to improve the effecti
veness of model fine-tuning. Moreover, we design a prompt pruning procedure to s
ystematically prune low importance prompts while preserving model performance, w
hich largely enhances the model's efficiency. Empirical results demonstrate that
 our approach outperforms several state-of-the-art baselines on two benchmarks,
with considerably low parameter usage (e.g., 0.32% of model parameters on VTAB-1

k). We anticipate that this work will inspire further exploration within the pre train-then-finetune paradigm for large-scale models.
****
High-Resolution Document Shadow Removal via A Large-Scale Real-World Dataset and A Frequency-Aware Shadow Erasing Net
Zinuo Li, Xuhang Chen, Chi-Man Pun, Xiaodong Cun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12449-12458
Shadows often occur when we capture the document with casual equipment, which influences the visual quality and readability of the digital copies. Different from the algorithms for natural shadow removal, the algorithms in document shadow removal need to preserve the details of fonts and figures in high-resolution input. Previous works ignore this problem and remove the shadows via approximate attention and small datasets, which might not work in real-world situations. We handle high-resolution document shadow removal directly via a larger-scale real-world dataset and a carefully-designed frequency-aware network. As for the dataset, we acquire over 7k couples of high-resolution (2462 x 3699) images of real-world documents pairs with various samples under different lighting circumstances, which is 10 times larger than existing datasets. As for the design of the network, we decouple the high-resolution images in the frequency domain, where the low-frequency details and high-frequency boundaries can be effectively learned via the carefully designed network structure. Powered by our network and dataset, the proposed method shows a clearly better performance than previous methods in terms of visual quality and numerical results. The code, models, and dataset are available at: https://github.com/CXH-Research/DocShadow-SD7K.
****
Scalable Diffusion Models with Transformers
William Peebles, Saining Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4195-4205
We explore a new class of diffusion models based on the transformer architecture. We train latent diffusion models of images, replacing the commonly-used U-Net backbone with a transformer that operates on latent patches. We analyze the scalability of our Diffusion Transformers (DiTs) through the lens of forward pass complexity as measured by Gflops. We find that DiTs with higher Gflops---through increased transformer depth/width or increased number of input tokens---consistently have lower FID. In addition to possessing good scalability properties, our largest DiT-XL/2 models outperform all prior diffusion models on the class-conditional ImageNet 512x512 and 256x256 benchmarks, achieving a state-of-the-art FID of 2.27 on the latter.
****
MMST-ViT: Climate Change-aware Crop Yield Prediction via Multi-Modal Spatial-Temporal Vision Transformer
Fudong Lin, Summer Crawford, Kaleb Guillot, Yihe Zhang, Yan Chen, Xu Yuan, Li Chen, Shelby Williams, Robert Minvielle, Xiangming Xiao, Drew Gholson, Nicolas Ashwell, Tri Setiyono, Brenda Tubana, Lu Peng, Magdy Bayoumi, Nian-Feng Tzeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5774-5784
Precise crop yield prediction provides valuable information for agricultural planning and decision-making processes. However, timely predicting crop yields remains challenging as crop growth is sensitive to growing season weather variation and climate change. In this work, we develop a deep learning-based solution, namely Multi-Modal Spatial-Temporal Vision Transformer (MMST-ViT), for predicting crop yields at the county level across the United States, by considering the effects of short-term meteorological variations during the growing season and the long-term climate change on crops. Specifically, our MMST-ViT consists of a Multi-Modal Transformer, a Spatial Transformer, and a Temporal Transformer. The Multi-Modal Transformer leverages both visual remote sensing data and short-term meteorological data for modeling the effect of growing season weather variations on crop growth. The Spatial Transformer learns the high-resolution spatial dependency among counties for accurate agricultural tracking. The Temporal Transformer captures the long-range temporal dependency for learning the impact of long-term c

limate change on crops. Meanwhile, we also devise a novel multi-modal contrastiv
e learning technique to pre-train our model without extensive human supervision.
 Hence, our MMST-ViT captures the impacts of both short-term weather variations
and long-term climate change on crops by leveraging both satellite images and me
teorological data. We have conducted extensive experiments on over 200 counties
in the United States, with the experimental results exhibiting that our MMST-ViT
 outperforms its counterparts under three performance metrics of interest.
****
From Knowledge Distillation to Self-Knowledge Distillation: A Unified Approach w
ith Normalized Loss and Customized Soft Labels
Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, Yu Li; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17
185-17194
Knowledge Distillation (KD) uses the teacher's prediction logits as soft labels
to guide the student, while self-KD does not need a real teacher to require the
soft labels. This work unifies the formulations of the two tasks by decomposing
and reorganizing the generic KD loss into a Normalized KD (NKD) loss and customi
zed soft labels for both target class (image's category) and non-target classes
named Universal Self-Knowledge Distillation (USKD). We decompose the KD loss and
 find the non-target loss from it forces the student's non-target logits to matc
h the teacher's, but the sum of the two non-target logits is different, preventi
ng them from being identical. NKD normalizes the non-target logits to equalize t
heir sum. It can be generally used for KD and self-KD to better use the soft lab
els for distillation loss. USKD generates customized soft labels for both target
 and non-target classes without a teacher. It smooths the target logit of the st
udent as the soft target label and uses the rank of the intermediate feature to
generate the soft non-target labels with Zipf's law. For KD with teachers, our N
KD achieves state-of-the-art performance on CIFAR-100 and ImageNet datasets, boo
sting the ImageNet Top-1 accuracy of ResNet18 from 69.90% to 71.96% with a ResNe
t-34 teacher. For self-KD without teachers, USKD is the first self-KD method tha
t can be effectively applied to both CNN and ViT models with negligible addition
al time and memory cost, resulting in new state-of-the-art results, such as 1.17
% and 0.55% accuracy gains on ImageNet for MobileNet and DeiT-Tiny, respectively
. Code is available at https://github.com/yzd-v/cls_KD.
****
SILT: Shadow-Aware Iterative Label Tuning for Learning to Detect Shadows from No
isy Labels
Han Yang, Tianyu Wang, Xiaowei Hu, Chi-Wing Fu; Proceedings of the IEEE/CVF Inte
rnational Conference on Computer Vision (ICCV), 2023, pp. 12687-12698
Existing shadow detection datasets often contain missing or mislabeled shadows,
which can hinder the performance of deep learning models trained directly on suc
h data. To address this issue, we propose SILT, the Shadow-aware Iterative Label
 Tuning framework, which explicitly considers noise in shadow labels and trains
the deep model in a self-training manner. Specifically, we incorporate strong da
ta augmentations with shadow counterfeiting to help the network better recognize
 non-shadow regions and alleviate overfitting. We also devise a simple yet effec
tive label tuning strategy with global-local fusion and shadow-aware filtering t
o encourage the network to make significant refinements on the noisy labels. We
evaluate the performance of SILT by relabeling the test set of the SBU dataset a
nd conducting various experiments. Our results show that even a simple U-Net tra
ined with SILT can outperform all state-of-the-art methods by a large margin. Wh
en trained on SBU / UCF / ISTD, our network can successfully reduce the Balanced
 Error Rate by 25.2% / 36.9% / 21.3% over the best state-of-the-art method.
****
Implicit Autoencoder for Point-Cloud Self-Supervised Representation Learning
Siming Yan, Zhenpei Yang, Haoxiang Li, Chen Song, Li Guan, Hao Kang, Gang Hua, Q
ixing Huang; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2023, pp. 14530-14542
This paper advocates the use of implicit surface representation in autoencoder-b
ased self-supervised 3D representation learning. The most popular and accessible

3D representation, i.e., point clouds, involves discrete samples of the underlying continuous 3D surface. This discretization process introduces sampling variations on the 3D shape, making it challenging to develop transferable knowledge of the true 3D geometry. In the standard autoencoding paradigm, the encoder is compelled to encode not only the 3D geometry but also information on the specific discrete sampling of the 3D shape into the latent code. This is because the point cloud reconstructed by the decoder is considered unacceptable unless there is a perfect mapping between the original and the reconstructed point clouds. This paper introduces the Implicit AutoEncoder (IAE), a simple yet effective method that addresses the sampling variation issue by replacing the commonly-used point-cloud decoder with an implicit decoder. The implicit decoder reconstructs a continuous representation of the 3D shape, independent of the imperfections in the discrete samples. Extensive experiments demonstrate that the proposed IAE achieves state-of-the-art performance across various self-supervised learning benchmarks.
****
Grounded Image Text Matching with Mismatched Relation Reasoning
Yu Wu, Yana Wei, Haozhe Wang, Yongfei Liu, Sibei Yang, Xuming He; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2976-2987
This paper introduces Grounded Image Text Matching with Mismatched Relation (GITM-MR), a novel visual-linguistic joint task that evaluates the relation understanding capabilities of transformer-based pre-trained models. GITM-MR requires a model to first determine if an expression describes an image, then localize referred objects or ground the mismatched parts of the text. We provide a benchmark for evaluating vision-language (VL) models on this task, with a focus on the challenging settings of limited training data and out-of-distribution sentence lengths. Our evaluation demonstrates that pre-trained VL models often lack data efficiency and length generalization ability. To address this, we propose the Relation-sensitive Correspondence Reasoning Network (RCRN), which incorporates relation-aware reasoning via bi-directional message propagation guided by language structure. Our RCRN can be interpreted as a modular program and delivers strong performance in terms of both length generalization and data efficiency. The code and data are available on https://github.com/SHTUPLUS/GITM-MR.
****
UniKD: Universal Knowledge Distillation for Mimicking Homogeneous or Heterogeneous Object Detectors
Shanshan Lao, Guanglu Song, Boxiao Liu, Yu Liu, Yujiu Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6362-6372
Knowledge distillation (KD) has become a standard method to boost the performance of lightweight object detectors. Most previous works are feature-based, where students mimic the features of homogeneous teacher detectors. However, distilling the knowledge from the heterogeneous teacher fails in this manner due to the serious semantic gap, which greatly limits the flexibility of KD in practical applications. Bridging this semantic gap now requires case-by-case algorithm design which is time-consuming and heavily relies on experienced adjustment. To alleviate this problem, we propose Universal Knowledge Distillation (UniKD), introducing additional decoder heads with deformable cross-attention called Adaptive Knowledge Extractor (AKE). In UniKD, AKEs are first pretrained on the teacher's output to infuse the teacher's content and positional knowledge into a fixed-number set of knowledge embeddings. The fixed AKEs are then attached to the student's backbone to encourage the student to absorb the teacher's knowledge in these knowledge embeddings. In this query-based distillation paradigm, detection-relevant information can be dynamically aggregated into a knowledge embedding set and transferred between different detectors. When the teacher model is too large for online inference, its output can be stored on disk in advance to save the computation overhead, which is more storage efficient than feature-based methods. Extensive experiments demonstrate that our UniKD can plug and play in any homogeneous or heterogeneous teacher-student pairs and significantly outperforms conventional feature-based KD.

****
Speech4Mesh: Speech-Assisted Monocular 3D Facial Reconstruction for Speech-Drive
n 3D Facial Animation
Shan He, Haonan He, Shuo Yang, Xiaoyan Wu, Pengcheng Xia, Bing Yin, Cong Liu, Li
rong Dai, Chang Xu; Proceedings of the IEEE/CVF International Conference on Comp
uter Vision (ICCV), 2023, pp. 14192-14202
Recent audio2mesh-based methods have shown promising prospects for speech-driven
 3D facial animation tasks. However, some intractable challenges are urgent to b
e settled. For example, the data-scarcity problem is intrinsically inevitable du
e to the difficulty of 4D data collection. Besides, current methods generally la
ck controllability on the animated face. To this end, we propose a novel framewo
rk named Speech4Mesh to consecutively generate 4D talking head data and train th
e audio2mesh network with the reconstructed meshes. In our framework, we first r
econstruct the 4D talking head sequence based on the monocular videos. For preci
se capture of the talking-related variation on the face, we exploit the audio-vi
sual alignment information from the video by employing a contrastive learning sc
heme. We next can train the audio2mesh network (e.g., FaceFormer) based on the g
enerated 4D data. To get control of the animated talking face, we encode the spe
aking-unrelated factors (e.g., emotion, etc.) into an emotion embedding for mani
pulation. Finally, a differentiable renderer guarantees more accurate photometri
c details of the reconstruction and animation results. Empirical experiments dem
onstrate that the Speech4Mesh framework can not only outperform state-of-the-art
 reconstruction methods, especially on the lower-face part but also achieve bett
er animation performance both perceptually and objectively after pre-trained on
the synthesized data. Besides, we also verify that the proposed framework is abl
e to explicitly control the emotion of the animated talking face.
****
BoxDiff: Text-to-Image Synthesis with Training-Free Box-Constrained Diffusion
Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng,
Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 7452-7461
Recent text-to-image diffusion models have demonstrated an astonishing capacity
to generate high-quality images. However, researchers mainly studied the way of
synthesizing images with only text prompts. While some works have explored using
 other modalities as conditions, considerable paired data, e.g., box/mask-image
pairs, and fine-tuning time are required for nurturing models. As such paired da
ta is time-consuming and labor-intensive to acquire and restricted to a closed s
et, this potentially becomes the bottleneck for applications in an open world. T
his paper focuses on the simplest form of user-provided conditions, e.g., box or
 scribble. To mitigate the aforementioned problem, we propose a training-free me
thod to control objects and contexts in the synthesized images adhering to the g
iven spatial conditions. Specifically, three spatial constraints, i.e., Inner-Bo
x, Outer-Box, and Corner Constraints, are designed and seamlessly integrated int
o the denoising step of diffusion models, requiring no additional training and m
assive annotated layout data. Extensive experimental results demonstrate that th
e proposed constraints can control what and where to present in the images while
 retaining the ability of Diffusion models to synthesize with high fidelity and
diverse concept coverage. The code is publicly available at https://github.com/s
howlab/BoxDiff.
****
Generalizing Neural Human Fitting to Unseen Poses With Articulated SE(3) Equivar
iance
Haiwen Feng, Peter Kulits, Shichen Liu, Michael J. Black, Victoria Fernandez Abr
evaya; Proceedings of the IEEE/CVF International Conference on Computer Vision (
ICCV), 2023, pp. 7977-7988
We address the problem of fitting a parametric human body model (SMPL) to point
cloud data. Optimization based methods require careful initialization and are pr
one to becoming trapped in local optima. Learning-based methods address this but
 do not generalize well when the input pose is far from those seen during traini
ng. For rigid point clouds, remarkable generalization has been achieved by lever

aging SE(3)-equivariant networks, but these methods do not work on articulated o
bjects. In this work we extend this idea to human bodies and propose ArtEq, a no
vel part-based SE(3)-equivariant neural architecture for SMPL model estimation f
rom point clouds. Specifically, we learn a part detection network by leveraging
local SO(3) invariance, and regress shape and pose using articulated SE(3) shape
-invariant and pose-equivariant networks, all trained end-to-end. Our novel pose
 regression module leverages the permutation-equivariant property of self-attent
ion layers to preserve rotational equivariance. Experimental results show that A
rtEq generalizes to poses not seen during training, outperforming state-of-the-a
rt methods by 44%in terms of body reconstruction accuracy, without requiring an
 optimization refinement step. Furthermore, ArtEq is three orders of magnitude f
aster during inference than prior work and has 97.3% fewer parameters. The code
and model are available for research purposes at https://arteq.is.tue.mpg.de.
****
Rapid Network Adaptation: Learning to Adapt Neural Networks Using Test-Time Feed
back
Teresa Yeo, O■uzhan Fatih Kar, Zahra Sodagar, Amir Zamir; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4674-4687
We propose a method for adapting neural networks to distribution shifts at test-
time. In contrast to training-time robustness mechanisms that attempt to anticip
ate the shift, we create a closed-loop system and make use of test-time feedback
 signal to adapt a network. We show that this loop can be effectively implemente
d using a learning-based function, which realizes an amortized optimizer for the
 network. This leads to an adaptation method, named Rapid Network Adaptation (RN
A), that is notably more flexible and orders of magnitude faster than the baseli
nes. Through a broad set of experiments using various adaptation signals and tar
get tasks, we study the generality, efficiency, and flexibility of this method.
We perform the evaluations using various datasets (Taskonomy, Replica, ScanNet,
Hypersim, COCO, ImageNet), tasks (depth, optical flow, semantic segmentation, cl
assification), and distribution shifts (Cross-datasets, 2D and 3D Common Corrupt
ions) with promising results.
****
Theoretical and Numerical Analysis of 3D Reconstruction Using Point and Line Inc
idences
Felix Rydell, Elima Shehu, Angélica Torres; Proceedings of the IEEE/CVF Internat
ional Conference on Computer Vision (ICCV), 2023, pp. 3748-3757
We study the joint image of lines incident to points, meaning the set of image t
uples obtained from fixed cameras observing a varying 3D point-line incidence. W
e prove a formula for the number of complex critical points of the triangulation
 problem that aims to compute a 3D point-line incidence from noisy images. Our f
ormula works for an arbitrary number of images and measures the intrinsic diffic
ulty of this triangulation. Additionally, we conduct numerical experiments using
 homotopy continuation methods, comparing different approaches of triangulation
of such incidences. In our setup, exploiting the incidence relations gives a not
ably faster point reconstruction with comparable accuracy.
****
Explaining Adversarial Robustness of Neural Networks from Clustering Effect Pers
pective
Yulin Jin, Xiaoyu Zhang, Jian Lou, Xu Ma, Zilong Wang, Xiaofeng Chen; Proceeding
s of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
4522-4531
Adversarial training (AT) is the most commonly used mechanism to improve the rob
ustness of deep neural networks. Recently, a novel adversarial attack against in
termediate layers exploits the extra fragility of adversarially trained networks
 to output incorrect predictions. The result implies the insufficiency in the se
arching space of the adversarial perturbation in adversarial training. To straig
hten out the reason for the effectiveness of the intermediate-layer attack, we i
nterpret the forward propagation as the Clustering Effect, characterizing that t
he intermediate-layer representations of neural networks for samples i.i.d. to t
he training set with the same label are similar, and we theoretically prove the

existence of Clustering Effect by corresponding Information Bottleneck Theory. We afterward observe that the intermediate-layer attack disobeys the clustering effect of the AT-trained model. Inspired by these significant observations, we propose a regularization method to extend the perturbation searching space during training, named sufficient adversarial training (SAT). We give a proven robustness bound of neural networks through rigorous mathematical proof. The experimental evaluations manifest the superiority of SAT over other state-of-the-art AT mechanisms in defending against adversarial attacks against both output and intermediate layers. Our code and Appendix can be found at https://github.com/clustering-effect/SAT.
****
Leaping Into Memories: Space-Time Deep Feature Synthesis
Alexandros Stergiou, Nikos Deligiannis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1966-1976
The success of deep learning models has led to their adaptation and adoption by prominent video understanding methods. The majority of these approaches encode features in a joint space-time modality for which the inner workings and learned representations are difficult to visually interpret. We propose LEArned Preconscious Synthesis (LEAPS), an architecture-independent method for synthesizing videos from the internal spatiotemporal representations of models. Using a stimulus video and a target class, we prime a fixed space-time model and iteratively optimize a video initialized with random noise. Additional regularizers are used to improve the feature diversity of the synthesized videos alongside the cross-frame temporal coherence of motions. We quantitatively and qualitatively evaluate the applicability of LEAPS by inverting a range of spatiotemporal convolutional and attention-based architectures trained on Kinetics-400, which to the best of our knowledge has not been previously accomplished.
****
Improving Generalization in Visual Reinforcement Learning via Conflict-aware Gradient Agreement Augmentation
Siao Liu, Zhaoyu Chen, Yang Liu, Yuzheng Wang, Dingkang Yang, Zhile Zhao, Ziqing Zhou, Xie Yi, Wei Li, Wenqiang Zhang, Zhongxue Gan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23436-23446
Learning a policy with great generalization to unseen environments remains challenging but critical in visual reinforcement learning. Despite the success of augmentation combination in the supervised learning generalization, naively applying it to visual RL algorithms may damage the training efficiency, suffering from serve performance degradation. In this paper, we first conduct qualitative analysis and illuminate the main causes: (i) high-variance gradient magnitudes and (ii) gradient conflicts existed in various augmentation methods. To alleviate these issues, we propose a general policy gradient optimization framework, named Conflict-aware Gradient Agreement Augmentation (CG2A), and better integrate augmentation combination into visual RL algorithms to address the generalization bias. In particular, CG2A develops a Gradient Agreement Solver to adaptively balance the varying gradient magnitudes, and introduces a Soft Gradient Surgery strategy to alleviate the gradient conflicts. Extensive experiments demonstrate that CG2A significantly improves the generalization performance and sample efficiency of visual RL algorithms.
****
Graph Matching with Bi-level Noisy Correspondence
Yijie Lin, Mouxing Yang, Jun Yu, Peng Hu, Changqing Zhang, Xi Peng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23362-23371
In this paper, we study a novel and widely existing problem in graph matching (GM), namely, Bi-level Noisy Correspondence (BNC), which refers to node-level noisy correspondence (NNC) and edge-level noisy correspondence (ENC). In brief, on the one hand, due to the poor recognizability and viewpoint differences between images, it is inevitable to inaccurately annotate some keypoints with offset and confusion, leading to the mismatch between two associated nodes, i.e., NNC. On the other hand, the noisy node-to-node correspondence will further contaminate th

e edge-to-edge correspondence, thus leading to ENC. For the BNC challenge, we propose a novel method termed Contrastive Matching with Momentum Distillation. Specifically, the proposed method is with a robust quadratic contrastive loss which enjoys the following merits: i) better exploring the node-to-node and edge-to-edge correlations through a GM customized quadratic contrastive learning paradigm; ii) adaptively penalizing the noisy assignments based on the confidence estimated by the momentum teacher. Extensive experiments on three real-world datasets show the robustness of our model compared with 12 competitive baselines. The code is available at https://github.com/XLearning-SCU/2023-ICCV-COMMON.
****
Learning from Noisy Pseudo Labels for Semi-Supervised Temporal Action Localization
Kun Xia, Le Wang, Sanping Zhou, Gang Hua, Wei Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10160-10169
Semi-Supervised Temporal Action Localization (SS-TAL) aims to improve the generalization ability of action detectors with large-scale unlabeled videos. Albeit the recent advancement, one of the major challenges still remains: noisy pseudo labels hinder efficient learning on abundant unlabeled videos, embodied as location biases and category errors. In this paper, we dive deep into such an important but understudied dilemma. To this end, we propose a unified framework, termed Noisy Pseudo-Label Learning, to handle both location biases and category errors. Specifically, our method is featured with (1) Noisy Label Ranking to rank pseudo labels based on the semantic confidence and boundary reliability, (2) Noisy Label Filtering to address the class-imbalance problem of pseudo labels caused by category errors, (3) Noisy Label Learning to penalize inconsistent boundary predictions to achieve noise-tolerant learning for heavy location biases. As a result, our method could effectively handle the label noise problem and improve the utilization of a large amount of unlabeled videos. Extensive experiments on THUMOS14 and ActivityNet v1.3 demonstrate the effectiveness of our method. The code is available at github.com/kunnxia/NPL.
****
InfiniCity: Infinite-Scale City Synthesis
Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, Sergey Tulyakov; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22808-22818
Toward infinite-scale 3D city synthesis, we propose a novel framework, InfiniCity, which constructs and renders an unconstrainedly large and 3D-grounded environment from random noises. InfiniCity decomposes the seemingly impractical task into three feasible modules, taking advantage of both 2D and 3D data. First, an infinite-pixel image synthesis module generates arbitrary-scale 2D maps from the bird's-eye view. Next, an octree-based voxel completion module lifts the generated 2D map to 3D octrees. Finally, a voxel-based neural rendering module texturizes the voxels and renders 2D images. InfiniCity can thus synthesize arbitrary-scale and traversable 3D city environments. We quantitatively and qualitatively demonstrate the efficacy of the proposed framework.
****
OpenOccupancy: A Large Scale Benchmark for Surrounding Semantic Occupancy Perception
Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, Xingang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17850-17859
Semantic occupancy perception is essential for autonomous driving, as automated vehicles require a fine-grained perception of the 3D urban structures. However, existing relevant benchmarks lack diversity in urban scenes, and they only evaluate front-view predictions. Towards a comprehensive benchmarking of surrounding perception algorithms, we propose OpenOccupancy, which is the first surrounding semantic occupancy perception benchmark. In the OpenOccupancy benchmark, we extend the large-scale nuScenes dataset with dense semantic occupancy annotations. Previous annotations rely on LiDAR points superimposition, where some occupancy labels are missed due to sparse LiDAR channels. To mitigate the problem, we intro

duce the Augmenting And Purifying (AAP) pipeline to 2x densify the annotations, where 4000 human hours are involved in the labeling process.
Besides, camera-based, LiDAR-based and multi-modal baselines are established for the OpenOccupancy benchmark. Furthermore, considering the complexity of surrounding occupancy perception lies in the computational burden of high-resolution 3D predictions, we propose the Cascade Occupancy Network (CONet) to refine the coarse prediction, which relatively enhances the performance by 30% than the baseline. We hope the OpenOccupancy benchmark will boost the development of surrounding occupancy perception algorithms.
****
Weakly-Supervised Text-Driven Contrastive Learning for Facial Behavior Understanding
Xiang Zhang, Taoyue Wang, Xiaotian Li, Huiyuan Yang, Lijun Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20751-20762
Contrastive learning has shown promising potential for learning robust representations by utilizing unlabeled data.
However, constructing effective positive-negative pairs for contrastive learning on facial behavior datasets remains challenging.
This is because such pairs inevitably encode the subject-ID information, and the randomly constructed pairs may push similar facial images away due to the limited number of subjects in facial behavior datasets.
To address this issue, we propose to utilize activity descriptions, coarse-grained information provided in some datasets, which can provide high-level semantic information about the image sequences but is often neglected in previous studies.
More specifically, we introduce a two-stage Contrastive Learning with Text-Embeded framework for Facial behavior understanding (CLEF).
The first stage is a weakly-supervised contrastive learning method that learns representations from positive-negative pairs constructed using coarse-grained activity information.
The second stage aims to train the recognition of facial expressions or facial action units by maximizing the similarity between image and the corresponding text label names.
The proposed CLEF achieves state-of-the-art performance on three in-the-lab datasets for AU recognition and three in-the-wild datasets for facial expression recognition.
****
Box-based Refinement for Weakly Supervised and Unsupervised Localization Tasks
Eyal Gomel, Tal Shaharbany, Lior Wolf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16044-16054
It has been established that training a box-based detector network can enhance the localization performance of weakly supervised and unsupervised methods. Moreover, we extend this understanding by demonstrating that these detectors can be utilized to improve the original network, paving the way for further advancements. To accomplish this, we train the detectors on top of the network output instead of the image data and apply suitable loss backpropagation. Our findings reveal a significant improvement in phrase grounding for the "what is where by looking" task, as well as various methods of unsupervised object discovery.
****
Activate and Reject: Towards Safe Domain Generalization under Category Shift
Chaoqi Chen, Luyao Tang, Leitian Tao, Hong-Yu Zhou, Yue Huang, Xiaoguang Han, Yizhou Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11552-11563
Albeit the notable performance on in-domain test points, it is non-trivial for deep neural networks to attain satisfactory accuracy when deploying in the open world, where novel domains and object classes often occur. In this paper, we study a practical problem of Domain Generalization under Category Shift (DGCS), which aims to simultaneously detect unknown-class samples and classify known-class samples in the target domains. Compared to prior DG works, we face two new challe

nges: 1) how to learn the concept of "unknown" during training with only source known-class samples, and 2) how to adapt the source-trained model to unseen environments for safe model deployment. To this end, we propose a novel Activate and Reject (ART) framework to reshape the model's decision boundary to accommodate unknown classes and conduct post hoc modification to further discriminate known and unknown classes using unlabeled test data. Specifically, during training, we promote the response to the unknown by optimizing the unknown probability and then smoothing the overall output to mitigate the overconfidence issue. At test time, we introduce a step-wise online adaptation method that predicts the label by virtue of the cross-domain nearest neighbor and class prototype information without updating the network's parameters or using threshold-based mechanisms. Experiments reveal that ART consistently improves the generalization capability of deep networks on different vision tasks. For image classification, ART improves the H-score by 6.1% on average compared to the previous best method. For object detection and semantic segmentation, we establish new benchmarks and achieve competitive performance.
****
PRIOR: Prototype Representation Joint Learning from Medical Images and Reports
Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, Xiaoying Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21361-21371
Contrastive learning based vision-language joint pre-training has emerged as a successful representation learning strategy. In this paper, we present a prototype representation learning framework incorporating both global and local alignment between medical images and reports. In contrast to standard global multi-modality alignment methods, we employ a local alignment module for fine-grained representation. Furthermore, a cross-modality conditional reconstruction module is designed to interchange information across modalities in the training phase by reconstructing masked images and reports. For reconstructing long reports, a sentence-wise prototype memory bank is constructed, enabling the network to focus on low-level localized visual and high-level clinical linguistic features. Additionally, a non-auto-regressive generation paradigm is proposed for reconstructing non-sequential reports. Experimental results on five downstream tasks, including supervised classification, zero-shot classification, image-to-text retrieval, semantic segmentation, and object detection, show the proposed method outperforms other state-of-the-art methods across multiple datasets and under different dataset size settings. The code is available at https://github.com/QtacierP/PRIOR.
****
Dynamic Mesh Recovery from Partial Point Cloud Sequence
Hojun Jang, Minkwan Kim, Jinseok Bae, Young Min Kim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15074-15084
The exact 3D dynamics of the human body provides crucial evidence to analyze the consequences of the physical interaction between the body and the environment, which can eventually assist everyday activities in a wide range of applications. However, optimizing for 3D configurations from image observation requires a significant amount of computation, whereas real-world 3D measurements often suffer from noisy observation or complex occlusion.
We resolve the challenge by learning a latent distribution representing strong temporal priors.
We use a conditional variational autoencoder (CVAE) architecture with a transformer to train the motion priors with a large-scale motion dataset.
Then our feature follower effectively aligns the feature spaces of noisy, partial observation with the necessary input for pre-trained motion priors, and quickly recovers a complete mesh sequence of motion.
We demonstrate that the transformer-based autoencoder can collect necessary spatio-temporal correlations robust to various adversaries, such as missing temporal frames, or noisy observation under severe occlusion.
Our framework is general and can be applied to recover the full 3D dynamics of other subjects with parametric representations.
****

WDiscOOD: Out-of-Distribution Detection via Whitened Linear Discriminant Analysis
Yiye Chen, Yunzhi Lin, Ruinian Xu, Patricio A. Vela; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5298-5307
Deep neural networks are susceptible to generating overconfident yet erroneous predictions when presented with data beyond known concepts. This challenge underscores the importance of detecting out-of-distribution (OOD) samples in the open world. In this work, we propose a novel feature-space OOD detection score based on class-specific and class-agnostic information. Specifically, the approach utilizes Whitened Linear Discriminant Analysis to project features into two subspaces - the discriminative and residual subspaces - for which the in-distribution (ID) classes are maximally separated and closely clustered, respectively. The OOD score is then determined by combining the deviation from the input data to the ID pattern in both subspaces. The efficacy of our method, named WDiscOOD, is verified on the large-scale ImageNet-1k benchmark, with six OOD datasets that cover a variety of distribution shifts. WDiscOOD demonstrates superior performance on deep classifiers with diverse backbone architectures, including CNN and vision transformer. Furthermore, we also show that WDiscOOD more effectively detects novel concepts in representation spaces trained with contrastive objectives, including supervised contrastive loss and multi-modality contrastive loss.
****
Boosting Few-shot Action Recognition with Graph-guided Hybrid Matching
Jiazheng Xing, Mengmeng Wang, Yudi Ruan, Bofan Chen, Yaowei Guo, Boyu Mu, Guang Dai, Jingdong Wang, Yong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1740-1750
Class prototype construction and matching are core aspects of few-shot action recognition. Previous methods mainly focus on designing spatiotemporal relation modeling modules or complex temporal alignment algorithms. Despite the promising results, they ignored the value of class prototype construction and matching, leading to unsatisfactory performance in recognizing similar categories in every task. In this paper, we propose GgHM, a new framework with Graph-guided Hybrid Matching. Concretely, we learn task-oriented features by the guidance of a graph neural network during class prototype construction, optimizing the intra- and inter-class feature correlation explicitly. Next, we design a hybrid matching strategy, combining frame-level and tuple-level matching to classify videos with multivariate styles. We additionally propose a learnable dense temporal modeling module to enhance the video feature temporal representation to build a more solid foundation for the matching process. GgHM shows consistent improvements over other challenging baselines on several few-shot datasets, demonstrating the effectiveness of our method. The code will be publicly available at https://github.com/jiazheng-xing/GgHM.
****
Neural Deformable Models for 3D Bi-Ventricular Heart Shape Reconstruction and Modeling from 2D Sparse Cardiac Magnetic Resonance Imaging
Meng Ye, Dong Yang, Mikael Kanski, Leon Axel, Dimitris Metaxas; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14247-14256
We propose a novel neural deformable model (NDM) targeting at the reconstruction and modeling of 3D bi-ventricular shape of the heart from 2D sparse cardiac magnetic resonance (CMR) imaging data. We model the bi-ventricular shape using blended deformable superquadrics, which are parameterized by a set of geometric parameter functions and are capable of deforming globally and locally. While global geometric parameter functions and deformations capture gross shape features from visual data, local deformations, parameterized as neural diffeomorphic point flows, can be learned to recover the detailed heart shape. Different from iterative optimization methods used in conventional deformable model formulations, NDMs can be trained to learn such geometric parameter functions, global and local deformations from a shape distribution manifold. Our NDM can learn to densify a sparse cardiac point cloud with arbitrary scales and generate high-quality triangular meshes automatically. It also enables the implicit learning of dense correspo

ndences among different heart shape instances for accurate cardiac shape registr
ation. Furthermore, the parameters of NDM are intuitive, and can be used by a ph
ysician without sophisticated post-processing. Experimental results on a large C
MR dataset demonstrate the improved performance of NDM over traditional methods.
****
Vision HGNN: An Image is More than a Graph of Nodes
Yan Han, Peihao Wang, Souvik Kundu, Ying Ding, Zhangyang Wang; Proceedings of th
e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19878-1
9888
The realm of graph-based modeling has proven its adaptability across diverse rea
l-world data types. However, its applicability to general computer vision tasks
had been limited until the introduction of the Vision Graph Neural Network (ViG)
. ViG divides input images into patches, conceptualized as nodes, constructing a
 graph through connections to nearest neighbors. Nonetheless, this method of gra
ph construction confines itself to simple pairwise relationships, leading to sur
plus edges and unwarranted memory and computation expenses. In this paper, we en
hance ViG by transcending conventional "pairwise" linkages and harnessing the po
wer of the hypergraph to encapsulate image information. Our objective is to enco
mpass more intricate inter-patch associations. In both training and inference ph
ases, we adeptly establish and update the hypergraph structure using the Fuzzy C
-Means method, ensuring minimal computational burden. This augmentation yields t
he Vision HyperGraph Neural Network (ViHGNN). The model's efficacy is empiricall
y substantiated through its state-of-the-art performance on both image classific
ation and object detection tasks, courtesy of the hypergraph structure learning
module that uncovers higher-order relationships. Our code is available at: https
://github.com/VITA-Group/ViHGNN.
****
Nonrigid Object Contact Estimation With Regional Unwrapping Transformer
Wei Xie, Zimeng Zhao, Shiying Li, Binghui Zuo, Yangang Wang; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9342-9351
Acquiring contact patterns between hands and nonrigid objects is a common concer
n in the vision and robotics community. However, existing learning-based methods
 focus more on contact with rigid ones from monocular images. When adopting them
 for nonrigid contact, a major problem is that the existing contact representati
on is restricted by the geometry of the object. Consequently, contact neighborho
ods are stored in an unordered manner and contact features are difficult to alig
n with image cues. At the core of our approach lies a novel hand-object contact
representation called RUPs (Region Unwrapping Profiles), which unwrap the roughl
y estimated hand-object surfaces as multiple high-resolution 2D regional profile
s. The region grouping strategy is consistent with the hand kinematic bone divis
ion because they are the primitive initiators for a composite contact pattern. B
ased on this representation, our Regional Unwrapping Transformer (RUFormer) lear
ns the correlation priors across regions from monocular inputs and predicts corr
esponding contact and deformed transformations. Our experiments demonstrate that
 the proposed framework can robustly estimate the deformed degrees and deformed
transformations, which make it suitable for both nonrigid and rigid contact.
****
Diffusion in Style
Martin Nicolas Everaert, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, Radhakrishn
a Achanta; Proceedings of the IEEE/CVF International Conference on Computer Visi
on (ICCV), 2023, pp. 2251-2261
We present Diffusion in Style, a simple method to adapt Stable Diffusion to any
desired style, using only a small set of target images. It is based on the key o
bservation that the style of the images generated by Stable Diffusion is tied to
 the initial latent tensor. Not adapting this initial latent tensor to the style
 makes fine-tuning slow, expensive, and impractical, especially when only a few
target style images are available. In contrast, fine-tuning is much easier if th
is initial latent tensor is also adapted. Our Diffusion in Style is orders of ma
gnitude more sample-efficient and faster. It also generates more pleasing images
 than existing approaches, as shown qualitatively and with quantitative comparis

ons.
****
FunnyBirds: A Synthetic Vision Dataset for a Part-Based Analysis of Explainable AI Methods
Robin Hesse, Simone Schaub-Meyer, Stefan Roth; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3981-3991
The field of explainable artificial intelligence (XAI) aims to uncover the inner workings of complex deep neural models. While being crucial for safety-critical domains, XAI inherently lacks ground-truth explanations, making its automatic evaluation an unsolved problem. We address this challenge by proposing a novel synthetic vision dataset, named FunnyBirds, and accompanying automatic evaluation protocols. Our dataset allows performing semantically meaningful image interventions, e.g., removing individual object parts, which has three important implications. First, it enables analyzing explanations on a part level, which is closer to human comprehension than existing methods that evaluate on a pixel level. Second, by comparing the model output for inputs with removed parts, we can estimate ground-truth part importances that should be reflected in the explanations. Third, by mapping individual explanations into a common space of part importances, we can analyze a variety of different explanation types in a single common framework. Using our tools, we report results for 24 different combinations of neural models and XAI methods, demonstrating the strengths and weaknesses of the assessed methods in a fully automatic and systematic manner.
****
Deformable Neural Radiance Fields using RGB and Event Cameras
Qi Ma, Danda Pani Paudel, Ajad Chhatkuli, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3590-3600
Modeling Neural Radiance Fields for fast-moving deformable objects from visual data alone is a challenging problem. A major issue arises due to the high deformation and low acquisition rates. To address this problem, we propose to use event cameras that offer very fast acquisition of visual change in an asynchronous manner. In this work, we develop a novel method to model the deformable neural radiance fields using RGB and Event cameras. The proposed method uses the asynchronous stream of events and calibrated sparse RGB frames. In this setup, the pose of the individual events --required to integrate them into the radiance fields-- remains to be unknown. Our method jointly optimizes the pose and the radiance field, in an efficient manner by leveraging the collection of events at once and actively sampling the events during learning. Experiments conducted on both realistically rendered and real-world datasets demonstrate a significant benefit of the proposed method over the state-of-the-art and the compared baseline. This shows a promising direction for modeling deformable neural radiance fields in real-world dynamic scenes. Our code and data will be publicly available.
****
BeLFusion: Latent Diffusion for Behavior-Driven Human Motion Prediction
German Barquero, Sergio Escalera, Cristina Palmero; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2317-2327
Stochastic human motion prediction (HMP) has generally been tackled with generative adversarial networks and variational autoencoders. Most prior works aim at predicting highly diverse motion in terms of the skeleton joints' dispersion. This has led to methods predicting fast and divergent movements, which are often unrealistic and incoherent with past motion. Such methods also neglect scenarios where anticipating diverse short-range behaviors with subtle joint displacements is important. To address these issues, we present BeLFusion, a model that, for the first time, leverages latent diffusion models in HMP to sample from a behavioral latent space where behavior is disentangled from pose and motion. Thanks to our behavior coupler, which is able to transfer sampled behavior to ongoing motion, BeLFusion's predictions display a variety of behaviors that are significantly more realistic, and coherent with past motion than the state of the art. To support it, we introduce two metrics, the Area of the Cumulative Motion Distribution, and the Average Pairwise Distance Error, which are correlated to realism according to a qualitative study (126 participants). Finally, we prove BeLFusion's

generalization power in a new cross-dataset scenario for stochastic HMP.
****
Semi-supervised Semantics-guided Adversarial Training for Robust Trajectory Prediction
Ruochen Jiao, Xiangguo Liu, Takami Sato, Qi Alfred Chen, Qi Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8207-8217
Predicting the trajectories of surrounding objects is a critical task for self-driving vehicles and many other autonomous systems. Recent works demonstrate that adversarial attacks on trajectory prediction, where small crafted perturbations are introduced to history trajectories, may significantly mislead the prediction of future trajectories and induce unsafe planning. However, few works have addressed enhancing the robustness of this important safety-critical task. In this paper, we present a novel adversarial training method for trajectory prediction. Compared with typical adversarial training on image tasks, our work is challenged by more random input with rich context and a lack of class labels. To address these challenges, we propose a method based on a semi-supervised adversarial autoencoder, which models disentangled semantic features with domain knowledge and provides additional latent labels for the adversarial training. Extensive experiments with different types of attacks demonstrate that our Semisupervised Semantics-guided Adversarial Training (SSAT) method can effectively mitigate the impact of adversarial attacks by up to 73% and outperform other popular defense methods. In addition, experiments show that our method can significantly improve the system's robust generalization to unseen patterns of attacks. We believe that such semantics-guided architecture and advancement on robust generalization is an important step for developing robust prediction models and enabling safe decision-making.
****
Linear-Covariance Loss for End-to-End Learning of 6D Pose Estimation
Fulin Liu, Yinlin Hu, Mathieu Salzmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14107-14117
Most modern image-based 6D object pose estimation methods learn to predict 2D-3D correspondences, from which the pose can be obtained using a PnP solver. Because of the non-differentiable nature of common PnP solvers, these methods are supervised via the individual correspondences. To address this, several methods have designed differentiable PnP strategies, thus imposing supervision on the pose obtained after the PnP step. Here, we argue that this conflicts with the averaging nature of the PnP problem, leading to gradients that may encourage the network to degrade the accuracy of individual correspondences. To address this, we derive a loss function that exploits the ground truth pose before solving the PnP problem. Specifically, we linearize the PnP solver around the ground-truth pose and compute the covariance of the resulting pose distribution. We then define our loss based on the diagonal covariance elements, which entails considering the final pose estimate yet not suffering from the PnP averaging issue. Our experiments show that our loss consistently improves the pose estimation accuracy for both dense and sparse correspondence based methods, achieving state-of-the-art results on both Linemod-Occluded and YCB-Video.
****
RLSAC: Reinforcement Learning Enhanced Sample Consensus for End-to-End Robust Estimation
Chang Nie, Guangming Wang, Zhe Liu, Luca Cavalli, Marc Pollefeys, Hesheng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9891-9900
Robust estimation is a crucial and still challenging task, which involves estimating model parameters in noisy environments. Although conventional sampling consensus-based algorithms sample several times to achieve robustness, these algorithms cannot use data features and historical information effectively. In this paper, we propose RLSAC, a novel Reinforcement Learning enhanced SAmple Consensus framework for end-to-end robust estimation. RLSAC employs a graph neural network to utilize both data and memory features to guide exploring directions for sampl

ing the next minimum set. The feedback of downstream tasks serves as the reward for unsupervised training. Therefore, RLSAC can avoid differentiating to learn the features and the feedback of downstream tasks for end-to-end robust estimation. In addition, RLSAC integrates a state transition module that encodes both data and memory features. Our experimental results demonstrate that RLSAC can learn from features to gradually explore a better hypothesis. Through analysis, it is apparent that RLSAC can be easily transferred to other sampling consensus-based robust estimation tasks. To the best of our knowledge, RLSAC is also the first method that uses reinforcement learning to sample consensus for end-to-end robust estimation. We
release our codes at https://github.com/IRMVLab/RLSAC.
****
CleanCLIP: Mitigating Data Poisoning Attacks in Multimodal Contrastive Learning
Hritik Bansal, Nishad Singhi, Yu Yang, Fan Yin, Aditya Grover, Kai-Wei Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 112-123
Multimodal contrastive pretraining has been used to train multimodal representation models, such as CLIP, on large amounts of paired image-text data. However, previous studies have revealed that such models are vulnerable to backdoor attacks. Specifically, when trained on backdoored examples, CLIP learns spurious correlations between the embedded backdoor trigger and the target label, aligning their representations in the joint embedding space. Injecting even a small number of poisoned examples, such as 75 examples in 3 million pretraining data, can significantly manipulate the model's behavior, making it difficult to detect or unlearn such correlations. To address this issue, we propose CleanCLIP, a finetuning framework that weakens the learned spurious associations introduced by backdoor attacks by independently re-aligning the representations for individual modalities. We demonstrate that unsupervised finetuning using a combination of multimodal contrastive and unimodal self-supervised objectives for individual modalities can significantly reduce the impact of the backdoor attack. Additionally, we show that supervised finetuning on task-specific labeled image data removes the backdoor trigger from the CLIP vision encoder. We show empirically that CleanCLIP maintains model performance on benign examples while erasing a range of backdoor attacks on multimodal contrastive learning.
****
Multi-Frequency Representation Enhancement with Privilege Information for Video Super-Resolution
Fei Li, Linfeng Zhang, Zikun Liu, Juan Lei, Zhenbo Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12814-12825
CNN's limited receptive field restricts its ability to capture long-range spatial-temporal dependencies, leading to unsatisfactory performance in video super-resolution. To tackle this challenge, this paper presents a novel multi-frequency representation enhancement module (MFE) that performs spatial-temporal information aggregation in the frequency domain. Specifically, MFE mainly includes a spatial-frequency representation enhancement branch which captures the long-range dependency in the spatial dimension, and an energy frequency representation enhancement branch to obtain the inter-channel feature relationship. Moreover, a novel model training method named privilege training is proposed to encode the privilege information from high-resolution videos to facilitate model training. With these two methods, we introduce a new VSR model named MFPI, which outperforms state-of-the-art methods by a large margin while maintaining good efficiency on various datasets, including REDS4, Vimeo, Vid4, and UDM10.
****
Self-supervised Pre-training for Mirror Detection
Jiaying Lin, Rynson W.H. Lau; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12227-12236
Existing mirror detection methods require supervised ImageNet pre-training to obtain good general-purpose image features. However, supervised ImageNet pre-training focuses on category-level discrimination and may not be suitable for downstream tasks like mirror detection, due to the overfitting upstream tasks (e.g., su

pervised image classification). We observe that mirror reflection is crucial to how people perceive the presence of mirrors, and such mid-level features can be better transferred from self-supervised pre-trained models. Inspired by this observation, in this paper we aim to improve mirror detection methods by proposing a new self-supervised learning (SSL) pre-training framework for modeling the representation of mirror reflection progressively in the pre-training process.
Our framework consists of three pre-training stages at different levels:
1) an image-level pre-training stage to globally incorporate mirror reflection features into the pre-trained model;
2) a patch-level pre-training stage to spatially simulate and learn local mirror reflection from image patches; and
3) a pixel-level pre-training stage to pixel-wisely capture mirror reflection via reconstructing corrupted mirror images based on the relationship between the inside and outside of mirrors.
Extensive experiments show that our SSL pre-training framework significantly outperforms previous state-of-the-art CNN-based SSL pre-training frameworks and even outperforms supervised ImageNet pre-training when transferred to the mirror detection task.
Code and models are available at https://jiaying.link/iccv2023-sslmirror/
****
GlowGAN: Unsupervised Learning of HDR Images from LDR Images in the Wild
Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, Thomas Leimkühler; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10509-10519
Most in-the-wild images are stored in Low Dynamic Range (LDR) form, serving as a partial observation of the High Dynamic Range (HDR) visual world. Despite limited dynamic range, these LDR images are often captured with different exposures, implicitly containing information about the underlying HDR image distribution. Inspired by this intuition, in this work we present, to the best of our knowledge, the first method for learning a generative model of HDR images from in-the-wild LDR image collections in a fully unsupervised manner. The key idea is to train a generative adversarial network (GAN) to generate HDR images which, when projected to LDR under various exposures, are indistinguishable from real LDR images. Experiments show that our method GlowGAN can synthesize photorealistic HDR images in many challenging cases such as landscapes, lightning, or windows, where previous supervised generative models produce overexposed images. With the assistance of GlowGAN, we showcase the innovative application of unsupervised inverse tone mapping (GlowGAN-ITM) that sets a new paradigm in this field. Unlike previous methods that gradually complete information from LDR input, GlowGAN-ITM searches the entire HDR image manifold modeled by GlowGAN for the HDR images which can be mapped back to the LDR input. GlowGAN-ITM method achieves more realistic reconstruction of overexposed regions compared to state-of-the-art supervised learning models, despite not requiring HDR images or paired multi-exposure images for training.
****
Cumulative Spatial Knowledge Distillation for Vision Transformers
Borui Zhao, Renjie Song, Jiajun Liang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6146-6155
Distilling knowledge from convolutional neural networks (CNNs) is a double-edged sword for vision transformers (ViTs). It boosts the performance since the image-friendly local-inductive bias of CNN helps ViT learn faster and better, but leading to two problems: (1) Network designs of CNN and ViT are completely different, which leads to different semantic levels of intermediate features, making spatial-wise knowledge transfer methods (e.g., feature mimicking) inefficient. (2) Distilling knowledge from CNN limits the network convergence in the later training period since ViT's capability of integrating global information is suppressed by CNN's local-inductive-bias supervision.
To this end, we present Cumulative Spatial Knowledge Distillation (CSKD). CSKD distills spatial-wise knowledge to all patch tokens of ViT from the corresponding spatial responses of CNN, without introducing intermediate features.

 Furthermore, CSKD exploits a Cumulative Knowledge Fusion (CKF) module, which introduces the global response of CNN and increasingly emphasizes its importance during the training. Applying CKF leverages CNN's local inductive bias in the early training period and gives full play to ViT's global capability in the later one. Extensive experiments and analysis on ImageNet-1k and downstream datasets demonstrate the superiority of our CSKD. Code will be publicly available.
****
Dual Pseudo-Labels Interactive Self-Training for Semi-Supervised Visible-Infrared Person Re-Identification
Jiangming Shi, Yachao Zhang, Xiangbo Yin, Yuan Xie, Zhizhong Zhang, Jianping Fan, Zhongchao Shi, Yanyun Qu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11218-11228
Visible-infrared person re-identification (VI-ReID) aims to match a specific person from a gallery of images captured from non-overlapping visible and infrared cameras. Most works focus on fully supervised VI-ReID, which requires substantial cross-modality annotation that is more expensive than the annotation in single-modality. To reduce the extensive cost of annotation, we explore two practical semi-supervised settings: uni-semi-supervised (annotating only visible images) and bi-semi-supervised (annotating partially in both modalities). These two semi-supervised settings face two challenges due to the large cross-modality discrepancies and the lack of correspondence supervision between visible and infrared images. Thus, it is difficult to generate reliable pseudo-labels and learn modality-invariant features from noise pseudo-labels. In this paper, we propose a dual pseudo-label interactive self-training (DPIS) for these two semi-supervised VI-ReID. Our DPIS integrates two pseudo-labels generated by distinct models into a hybrid pseudo-label for unlabeled data. However, the hybrid pseudo-label still inevitably contains noise. To eliminate the negative effect of noise pseudo-labels, we introduce three modules: noise label penalty (NLP), noise correspondence calibration (NCC), and unreliable anchor learning (UAL). Specifically, NLP penalizes noise labels, NCC calibrates noisy correspondences, and UAL mines the hard-to-discriminate features. Extensive experimental results on SYSU-MM01 and RegDB demonstrate that our DPIS achieves impressive performance under these two semi-supervised settings.
****
Less is More: Focus Attention for Efficient DETR
Dehua Zheng, Wenhui Dong, Hailin Hu, Xinghao Chen, Yunhe Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6674-6683
DETR-like models have significantly boosted the performance of detectors and even outperformed classical convolutional models. However, all tokens are treated equally without discrimination brings a redundant computational burden in the traditional encoder structure. The recent sparsification strategies exploit a subset of informative tokens to reduce attention
 complexity maintaining performance through the sparse encoder. But these methods tend to rely on unreliable model statistics. Moreover, simply reducing the token population hinders the detection performance to a large extent, limiting the application of these sparse models. We propose Focus-DETR, which focuses attention on more informative tokens for a better trade-off between computation efficiency and model accuracy. Specifically, we reconstruct the encoder with dual attention, which includes a token scoring mechanism that considers both localization and category semantic information of the objects from multi-scale feature maps. We efficiently abandon the background queries and enhance the semantic interaction of the fine-grained object queries based on the scores. Compared with the state-of-the-art sparse DETR-like detectors under the same setting, our Focus-DETR gets comparable complexity while achieving 50.4AP (+2.2) on COCO. The code is available at https://github.com/huawei-noah/noah-research/tree/master/Focus-DETR and https://gitee.com/mindspore/models/tree/master/research/cv/Focus-DETR.
****
Efficient Controllable Multi-Task Architectures
Abhishek Aich, Samuel Schulter, Amit K. Roy-Chowdhury, Manmohan Chandraker, Yumi

n Suh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5740-5751
We aim to train a multi-task model such that users can adjust the desired compute budget and relative importance of task performances after deployment, without retraining. This enables optimizing performance for dynamically varying user needs, without heavy computational overhead to train and save models for various scenarios. To this end, we propose a multi-task model consisting of a shared encoder and task-specific decoders where both encoder and decoder channel widths are slimmable. Our key idea is to control the task importance by varying the capacities of task-specific decoders, while controlling the total computational cost by jointly adjusting the encoder capacity. This improves overall accuracy by allowing a stronger encoder for a given budget, increases control over computational cost, and delivers high-quality slimmed sub-architectures based on user's constraints. Our training strategy involves a novel `Configuration-Invariant Knowledge Distillation' loss that enforces backbone representations to be invariant under different runtime width configurations to enhance accuracy. Further, we present a simple but effective search algorithm that translates user constraints to runtime width configurations of both the shared encoder and task decoders, for sampling the sub-architectures. The key rule for the search algorithm is to provide a larger computational budget to the higher preferred task decoder, while searching a shared encoder configuration that enhances the overall MTL performance. Various experiments on three multi-task benchmarks (PASCALContext, NYUDv2, and CIFAR100-MTL) with diverse backbone architectures demonstrate the advantage of our approach. For example, our method shows a higher controllability by 33.5% in the NYUD-v2 dataset over prior methods, while incurring much less compute cost.
****
HumanSD: A Native Skeleton-Guided Diffusion Model for Human Image Generation
Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, Qiang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15988-15998
Controllable human image generation (HIG) has attracted significant attention from academia and industry for its numerous real-life applications. State-of-the-art solutions, such as ControlNet and T2I-Adapter, introduce an additional learnable branch on top of the frozen pre-trained stable diffusion (SD) model, which can enforce various kinds of conditions, including skeleton guidance of HIG. While such a plug-and-play approach is appealing, the inevitable and uncertain conflicts between the original images produced from the frozen SD branch and the given condition incur significant challenges for the learnable branch, which conduct the condition learning via image feature editing.
In this work, we propose a native skeleton-guided diffusion model for controllable HIG called HumanSD. Instead of performing image editing with dual-branch diffusion, we fine-tune the original SD model using a novel heatmap-guided denoising loss. This strategy effectively and efficiently strengthens the given skeleton condition during model training while mitigating the catastrophic forgetting effects. HumanSD is fine-tuned on the assembly of three large-scale human-centric datasets with text-image-pose information, two of which are established in this work. Experimental results show that HumanSD outperforms ControlNet in terms of pose control and image quality, particularly when the given skeleton guidance is sophisticated. Code and data are available at: https://idearesearch.github.io/HumanSD/.
****
Lens Parameter Estimation for Realistic Depth of Field Modeling
Dominique Piché-Meunier, Yannick Hold-Geoffroy, Jianming Zhang, Jean-François Lalonde; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 499-508
We present a method to estimate the depth of field effect from a single image. Most existing methods related to this task provide either a per-pixel estimation of blur and/or depth. Instead, we go further and propose to use a lens-based representation that models the depth of field using two parameters: the blur factor and focus disparity. Those two parameters, along with the signed defocus repres

entation, result in a more intuitive and linear representation which we solve us ing a novel weighting network. Furthermore, our method explicitly enforces consi stency between the estimated defocus blur, the lens parameters, and the depth ma p. Finally, we train our deep-learning-based model on a mix of real images with synthetic depth of field and fully synthetic images. These improvements result i n a more robust and accurate method, as demonstrated by our state-of-the-art res ults. In particular, our lens parametrization enables several applications, such as 3D staging for AR environments and seamless object compositing.
****
Learned Compressive Representations for Single-Photon 3D Imaging
Felipe Gutierrez-Barragan, Fangzhou Mu, Andrei Ardelean, Atul Ingle, Claudio Bru schini, Edoardo Charbon, Yin Li, Mohit Gupta, Andreas Velten; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10756-10 766
Single-photon 3D cameras can record the time-of-arrival of billions of photons p er second with picosecond accuracy. One common approach to summarize the photon data stream is to build a per-pixel timestamp histogram, resulting in a 3D histo gram tensor that encodes distances along the time axis. As the spatio-temporal r esolution of the histogram tensor increases, the in-pixel memory requirements an d output data rates can quickly become impractical. To overcome this limitation, we propose a family of linear compressive representations of histogram tensors that can be computed efficiently, in an online fashion, as a matrix operation. W e design practical lightweight compressive representations that are amenable to an in-pixel implementation and consider the spatio-temporal information of each timestamp. Furthermore, we implement our proposed framework as the first layer o f a neural network, which enables the joint end-to-end optimization of the compr essive representations and a downstream SPAD data processing model. We find that a well-designed compressive representation can reduce in-sensor memory and data rates up to 2 orders of magnitude without significantly reducing 3D imaging qua lity. Finally, we analyze the power consumption implications through an on-chip implementation.
****
Alignment-free HDR Deghosting with Semantics Consistent Transformer
Steven Tel, Zongwei Wu, Yulun Zhang, Barthélémy Heyrman, Cédric Demonceaux, Radu Timofte, Dominique Ginhac; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12836-12845
High dynamic range (HDR) imaging aims to retrieve information from multiple low- dynamic range inputs to generate realistic output. The essence is to leverage th e contextual information, including both dynamic and static semantics, for bette r image generation. Existing methods often focus on the spatial misalignment acr oss input frames caused by the foreground and/or camera motion. However, there i s no research on jointly leveraging the dynamic and static context in a simultan eous manner. To delve into this problem, we propose a novel alignment-free netwo rk with a Semantics Consistent Transformer (SCTNet) with both spatial and channe l attention modules in the network. The spatial attention aims to deal with the intra-image correlation to model the dynamic motion, while the channel attention enables the inter-image intertwining to enhance the semantic consistency across frames. Aside from this, we introduce a novel realistic HDR dataset with more v ariations in foreground objects, environmental factors, and larger motions. Exte nsive comparisons on both conventional datasets and ours validate the effectiven ess of our method, achieving the best trade-off on the performance and the compu tational cost. The source code and dataset are available at https://github.com/Z ongwei97/SCTNet.
****
Semantic-Aware Implicit Template Learning via Part Deformation Consistency
Sihyeon Kim, Minseok Joo, Jaewon Lee, Juyeon Ko, Juhan Cha, Hyunwoo J. Kim; Proc eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023 , pp. 593-603
Learning implicit templates as neural fields has recently shown impressive perfo rmance in unsupervised shape correspondence. Despite the success, we observe cur

rent approaches, which solely rely on geometric information, often learn subopti
mal deformation across generic object shapes, which have high structural variabi
lity. In this paper, we highlight the importance of part deformation consistency
 and propose a semantic-aware implicit template learning framework to enable sem
antically plausible deformation. By leveraging semantic prior from a self-superv
ised feature extractor, we suggest local conditioning with novel semantic-aware
deformation code and deformation consistency regularizations regarding part defo
rmation, global deformation, and global scaling. Our extensive experiments demon
strate the superiority of the proposed method over baselines in various tasks: k
eypoint transfer, part label transfer, and texture transfer. More interestingly,
 our framework shows a larger performance gain under more challenging settings.
We also provide qualitative analyses to validate the effectiveness of semantic-a
ware deformation. The code is available at https://github.com/mlvlab/PDC.
****
HRS-Bench: Holistic, Reliable and Scalable Benchmark for Text-to-Image Models
Eslam Mohamed Bakr, Pengzhan Sun, Xiaogian Shen, Faizan Farooq Khan, Li Erran Li
, Mohamed Elhoseiny; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2023, pp. 20041-20053
Designing robust text-to-image (T2I) models have been extensively explored in re
cent years, especially with the emergence of diffusion models, which achieves st
ate-of-the-art results on T2I synthesis tasks. Despite the significant effort an
d success in this direction, we observed that the existing metrics need to be mo
re robust to measure real progress. Therefore, comparing the existing models are
 more complex and heavily subjective for human evaluations. In addition, we obse
rve that the efforts in developing new architectures do not coincide with effort
s in the evaluation direction. Driven by this observation, the importance of des
igning a concrete evaluation emerges to fill the gap between designing and evalu
ation efforts. Accordingly, we introduce our holistic, reliable, and scalable be
nchmark, termed \papernameAbbrev , for T2I models. Unlike the existing benchmar
ks, which focus on limited aspects, we measure 13 skills, which could be categor
ized into five critical skills; accuracy, robustness, generalization, fairness,
and bias. In addition, \papernameAbbrev covers 50 applications, e.g., fashion,
 animals, transportation, food, and clothes. We evaluate nine recent large-scale
 T2I models using metrics that cover a wide range of skills. We study 13 skills,
 e.g., robustness, fairness, and bias. To probe the effectiveness of our \paper
nameAbbrev , a human evaluation is conducted, which is aligned with 95% with our
 evaluations on average across the 13 skills. We hope our findings, e.g., all th
e existing models can not generate visual text nor emotionally grounded images,
help accelerate and direct future research. To this end, the code and data are a
vailable at https://eslambakr.github.io/hrsbench.github.io/.
****
Multi3DRefer: Grounding Text Description to Multiple 3D Objects
Yiming Zhang, ZeMing Gong, Angel X. Chang; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2023, pp. 15225-15236
We introduce the task of localizing a flexible number of objects in real-world 3
D scenes using natural language descriptions. Existing 3D visual grounding tasks
 focus on localizing a unique object given a text description. However, such a s
trict setting is unnatural as localizing potentially multiple objects is a commo
n need in real-world scenarios and robotic tasks (e.g., visual navigation and ob
ject rearrangement). To address this setting we propose Multi3DRefer, generalizi
ng the ScanRefer dataset and task. Our dataset contains 61926 descriptions of 11
609 objects, where zero, single or multiple target objects are referenced by eac
h description. We also introduce a new evaluation metric and benchmark methods f
rom prior work to enable further investigation of multi-modal 3D scene understan
ding. Furthermore, we develop a better baseline leveraging 2D features from CLIP
 by rendering object proposals online with contrastive learning, which outperfor
ms the state of the art on the ScanRefer benchmark.
****
Examining Autoexposure for Challenging Scenes
SaiKiran Tedla, Beixuan Yang, Michael S. Brown; Proceedings of the IEEE/CVF Inte

rnational Conference on Computer Vision (ICCV), 2023, pp. 13076-13085
Autoexposure (AE) is a critical step applied by camera systems to ensure properl
y exposed images. While current AE algorithms are effective in well-lit environm
ents with constant illumination, these algorithms still struggle in environments
 with bright light sources or scenes with abrupt changes in lighting. A signific
ant hurdle in developing new AE algorithms for challenging environments, especia
lly those with time-varying lighting, is the lack of suitable image datasets. To
 address this issue, we have captured a new 4D exposure dataset that provides a
large solution space (i.e., shutter speed range from 1/500 to 15 seconds) over a
 temporal sequence with moving objects, bright lights, and varying lighting. In
addition, we have designed a software platform to allow AE algorithms to be used
 in a plug-and-play manner with the dataset. Our dataset and associate platform
enable repeatable evaluation of different AE algorithms and provide a much-neede
d starting point to develop better AE methods. We examine several existing AE st
rategies using our dataset and show that most users prefer a simple saliency met
hod for challenging lighting conditions.
****
DiffCloth: Diffusion Based Garment Synthesis and Manipulation via Structural Cro
ss-modal Semantic Alignment
Xujie Zhang, Binbin Yang, Michael C. Kampffmeyer, Wenqing Zhang, Shiyue Zhang, G
uansong Lu, Liang Lin, Hang Xu, Xiaodan Liang; Proceedings of the IEEE/CVF Inter
national Conference on Computer Vision (ICCV), 2023, pp. 23154-23163
Cross-modal garment synthesis and manipulation will significantly benefit the wa
y fashion designers generate garments and modify their designs via flexible ling
uistic interfaces. However, despite the significant progress that has been made
in generic image synthesis using diffusion models, producing garment images with
 garment part level semantics that are well aligned with input text prompts and
then flexibly manipulating the generated results still remains a problem. Curren
t approaches follow the general text-to-image paradigm and mine cross-modal rela
tions via simple cross-attention modules, neglecting the structural corresponden
ce between visual and textual representations in the fashion design domain. In t
his work, we instead introduce DiffCloth, a diffusion-based pipeline for cross-m
odal garment synthesis and manipulation, which empowers diffusion models with fl
exible compositionality in the fashion domain by structurally aligning the cross
-modal semantics. Specifically, we formulate the part-level cross-modal alignmen
t as a bipartite matching problem between the linguistic Attribute-Phrases (AP)
and the visual garment parts which are obtained via constituency parsing and sem
antic segmentation, respectively. To mitigate the issue of attribute confusion,
we further propose a semantic-bundled cross-attention to preserve the spatial st
ructure similarities between the attention maps of attribute adjectives and part
 nouns in each AP. Moreover, DiffCloth allows for manipulation of the generated
results by simply replacing APs in the text prompts. The manipulation-irrelevant
 regions are recognized by blended masks obtained from the bundled attention map
s of the APs and kept unchanged. Extensive experiments on the CM-Fashion benchma
rk demonstrate that DiffCloth both yields state-of-the-art garment synthesis res
ults by leveraging the inherent structural information and supports flexible man
ipulation with region consistency.
****
Improved Visual Fine-tuning with Natural Language Supervision
Junyang Wang, Yuanhong Xu, Juhua Hu, Ming Yan, Jitao Sang, Qi Qian; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11
899-11909
Fine-tuning a visual pre-trained model can leverage the semantic information fro
m large-scale pre-training data and mitigate the over-fitting problem on downstr
eam vision tasks with limited training examples. While the problem of catastroph
ic forgetting in pre-trained backbone has been extensively studied for fine-tuni
ng, its potential bias from the corresponding pre-training task and data, attrac
ts less attention. In this work, we investigate this problem by demonstrating th
at the obtained classifier after fine-tuning will be close to that induced by th
e pre-trained model. To reduce the bias in the classifier effectively, we introd

uce a reference distribution obtained from a fixed text classifier, which can he lp regularize the learned vision classifier. The proposed method, Text Supervise d fine-tuning (TeS), is evaluated with diverse pre-trained vision models includi ng ResNet and ViT, and text encoders including BERT and CLIP, on 11 downstream t asks. The consistent improvement with a clear margin over distinct scenarios con firms the effectiveness of our proposal. Code is available at https://github.com /idstcv/TeS.
****
Person Re-Identification without Identification via Event anonymization
Shafiq Ahmad, Pietro Morerio, Alessio Del Bue; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2023, pp. 11132-11141
Wide-scale use of visual surveillance in public spaces puts individual privacy a t stake while increasing resource consumption (energy, bandwidth, and computatio n). Neuromorphic vision sensors (event-cameras) have been recently considered a valid solution to the privacy issue because they do not capture detailed RGB vis ual information of the subjects in the scene. However, recent deep learning arch itectures have been able to reconstruct images from event cameras with high fide lity, reintroducing a potential threat to privacy for event-based vision applica tions. In this paper, we aim to anonymize event-streams to protect the identity of human subjects against such image reconstruction attacks. To achieve this, we propose an end-to-end network architecture jointly optimized for the twofold ob jective of preserving privacy and performing a downstream task such as person Re Id. Our network learns to scramble events, enforcing the degradation of images r ecovered from the privacy attacker. In this work, we also bring to the community the first ever event-based person ReId dataset gathered to evaluate the perform ance of our approach. We validate our approach with extensive experiments and re port results on the synthetic event data simulated from the publicly available S oftBio dataset and our proposed Event-ReId dataset.
****
GRAM-HD: 3D-Consistent Image Generation at High Resolution with Generative Radia nce Manifolds
Jianfeng Xiang, Jiaolong Yang, Yu Deng, Xin Tong; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2023, pp. 2195-2205
Recent works have shown that 3D-aware GANs trained on unstructured single image collections can generate multiview images of novel instances. The key underpinni ngs to achieve this are a 3D radiance field generator and a volume rendering pro cess. However, existing methods either cannot generate high-resolution images (e .g., up to 256x256) due to the high computation cost of neural volume rendering, or rely on 2D CNNs for image-space upsampling which jeopardizes the 3D consiste ncy across different views. This paper proposes a novel 3D-aware GAN that can ge nerate high resolution images (up to 1024x1024) while keeping strict 3D consiste ncy as in volume rendering. Our motivation is to achieve super-resolution direct ly in the 3D space to preserve 3D consistency. We avoid the otherwise prohibitiv ely-expensive computation cost by applying 2D convolutions on a set of 2D radian ce manifolds defined in the recent generative radiance manifold (GRAM) approach, and apply dedicated loss functions for effective GAN training at high resolutio n. Experiments on FFHQ and AFHQv2 datasets show that our method can produce high -quality 3D-consistent results that significantly outperform existing methods. I t makes a significant step towards closing the gap between traditional 2D image generation and 3D-consistent free-view generation.
****
Small Object Detection via Coarse-to-fine Proposal Generation and Imitation Lear ning
Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, Junwei Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6317-632 7
The past few years have witnessed the immense success of object detection, while current excellent detectors struggle on tackling size-limited instances. Concre tely, the well-known challenge of low overlaps between the priors and object reg ions leads to a constrained sample pool for optimization, and the paucity of dis

criminative information further aggravates the recognition. To alleviate the aforementioned issues, we propose CFINet, a two-stage framework tailored for small object detection based on the Coarse-to-fine pipeline and Feature Imitation learning. Firstly, we introduce Coarse-to-fine RPN (CRPN) to ensure sufficient and high-quality proposals for small objects through the dynamic anchor selection strategy and cascade regression. Then, we equip the conventional detection head with a Feature Imitation (FI) branch to facilitate the region representations of size-limited instances that perplex the model in an imitation manner. Moreover, an auxiliary imitation loss following supervised contrastive learning paradigm is devised to optimize this branch. When integrated with Faster RCNN, CFINet achieves state-of-the-art performance on the large-scale small object detection benchmarks, SODA-D and SODA-A, underscoring its superiority over baseline detector and other mainstream detection approaches. Code is available at https://github.com/shaunyuan22/CFINet.
****
Anomaly Detection Under Distribution Shift
Tri Cao, Jiawen Zhu, Guansong Pang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6511-6523
Anomaly detection (AD) is a crucial machine learning task that aims to learn patterns from a set of normal training samples to identify abnormal samples in test data. Most existing AD studies assume that the training and test data are drawn from the same data distribution, but the test data can have large distribution shifts arising in many real-world applications due to different natural variations such as new lighting conditions, object poses, or background appearances, rendering existing AD methods ineffective in such cases. In this paper, we consider the problem of anomaly detection under distribution shift and establish performance benchmarks on four widely-used AD and out-of-distribution (OOD) generalization datasets. We demonstrate that simple adaptation of state-of-the-art OOD generalization methods to AD settings fails to work effectively due to the lack of labeled anomaly data. We further introduce a novel robust AD approach to diverse distribution shifts by minimizing the distribution gap between in-distribution and OOD normal samples in both the training and inference stages in an unsupervised way. Our extensive empirical results on the four datasets show that our approach substantially outperforms state-of-the-art AD methods and OOD generalization methods on data with various distribution shifts, while maintaining the detection accuracy on in-distribution data. Code and data are available at https://github.com/mala-lab/ADShift.
****
Class Prior-Free Positive-Unlabeled Learning with Taylor Variational Loss for Hyperspectral Remote Sensing Imagery
Hengwei Zhao, Xinyu Wang, Jingtao Li, Yanfei Zhong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16827-16836
Positive-unlabeled learning (PU learning) in hyperspectral remote sensing imagery (HSI) is aimed at learning a binary classifier from positive and unlabeled data, which has broad prospects in various earth vision applications. However, when PU learning meets limited labeled HSI, the unlabeled data may dominate the optimization process, which makes the neural networks overfit the unlabeled data. In this paper, a Taylor variational loss is proposed for HSI PU learning, which reduces the weight of the gradient of the unlabeled data by Taylor series expansion to enable the network to find a balance between overfitting and underfitting. In addition, the self-calibrated optimization strategy is designed to stabilize the training process. Experiments on 7 benchmark datasets (21 tasks in total) validate the effectiveness of the proposed method. Code is at: https://github.com/Hengwei-Zhao96/T-HOneCls.
****
HoloAssist: an Egocentric Human Interaction Dataset for Interactive AI Assistants in the Real World
Xin Wang, Taein Kwon, Mahdi Rad, Bowen Pan, Ishani Chakraborty, Sean Andrist, Dan Bohus, Ashley Feniello, Bugra Tekin, Felipe Vieira Frujeri, Neel Joshi, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Visio

n (ICCV), 2023, pp. 20270-20281
Building an interactive AI assistant that can perceive, reason, and collaborate with humans in the real world has been a long-standing pursuit in the AI community. This work is part of a broader research effort to develop intelligent agents that can interactively guide humans through performing tasks in the physical world. As a first step in this direction, we introduce HoloAssist, a large-scale egocentric human interaction dataset, where two people collaboratively complete physical manipulation tasks. The task performer executes the task while wearing a mixed-reality headset that captures seven synchronized data streams. The task instructor watches the performer's egocentric video in real time and guides them verbally. By augmenting the data with action and conversational annotations and observing the rich behaviors of various participants, we present key insights into how human assistants correct mistakes, intervene in the task completion procedure, and ground their instructions to the environment. HoloAssist spans 166 hours of data captured by 350 unique instructor-performer pairs. Furthermore, we construct and present benchmarks on mistake detection, intervention type prediction, and hand forecasting, along with detailed analysis. We expect HoloAssist will provide an important resource for building AI assistants that can fluidly collaborate with humans in the real world. Data can be downloaded at https://holoassist.github.io/.
****
Self-Feedback DETR for Temporal Action Detection
Jihwan Kim, Miso Lee, Jae-Pil Heo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10286-10296
Temporal Action Detection (TAD) is challenging but fundamental for real-world video applications. Recently, DETR-based models have been devised for TAD but have not performed well yet. In this paper, we point out the problem in the self-attention of DETR for TAD; the attention modules focus on a few key elements, called temporal collapse problem. It degrades the capability of the encoder and decoder since their self-attention modules play no role. To solve the problem, we propose a novel framework, Self-DETR, which utilizes cross-attention maps of the decoder to reactivate self-attention modules. We recover the relationship between encoder features by simple matrix multiplication of the cross-attention map and its transpose. Likewise, we also get the information within decoder queries. By guiding collapsed self-attention maps with the guidance map calculated, we settle down the temporal collapse of self-attention modules in the encoder and decoder. Our extensive experiments demonstrate that Self-DETR resolves the temporal collapse problem by keeping high diversity of attention over all layers.
****
StableVideo: Text-driven Consistency-aware Diffusion Video Editing
Wenhao Chai, Xun Guo, Gaoang Wang, Yan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23040-23050
Diffusion-based methods can generate realistic images and videos, but they struggle to edit existing objects in a video while preserving their geometry over time. This prevents diffusion models from being applied to natural video editing. In this paper, we tackle this problem by introducing temporal dependency to existing text-driven diffusion models, which allows them to generate consistent appearance for the new objects. Specifically, we develop a novel inter-frame propagation mechanism for diffusion video editing, which leverages the concept of layered representations to propagate the geometry and appearance information from one frame to the next. We then build up a text-driven video editing framework based on this mechanism, namely StableVideo, which can achieve consistency-aware video editing. Extensive qualitative experiments demonstrate the strong editing capability of our approach. Compared with state-of-the-art video editing methods, our approach shows superior qualitative and quantitative results.
****
PIRNet: Privacy-Preserving Image Restoration Network via Wavelet Lifting
Xin Deng, Chao Gao, Mai Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22368-22377
The cloud-based multimedia service becomes increasingly popular in the last deca

de, however, it poses a serious threat to the client's privacy. To address this issue, many methods utilized image encryption as a defense mechanism. However, the encrypted images look quite different from the natural images, making them vulnerable to attackers. In this paper, we propose a novel method namely PIRNet, which operates privacy-preserving image restoration in the steganographic domain. Compared to existing methods, our method offers significant advantages in terms of invisibility and security. Specifically, we first propose a wavelet Lifting-based Invertible Hiding (LIH) network to conceal the secret image into the stego image. Then, a Lifting-based Secure Restoration (LSR) network is utilized to perform image restoration in the steganographic domain. Since the secret image remains hidden throughout the whole image restoration process, the privacy of clients can be largely ensured. In addition, since the stego image looks visually the same as the cover image, the attackers can hardly discover it, which significantly improves the security. The experimental results on different datasets show the superiority of our PIRNet over the existing methods on various privacy-preserving image restoration tasks, including image denoising, deblurring and super-resolution.
****
LAW-Diffusion: Complex Scene Generation by Diffusion with Layouts
Binbin Yang, Yi Luo, Ziliang Chen, Guangrun Wang, Xiaodan Liang, Liang Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22669-22679
Thanks to the rapid development of diffusion models, unprecedented progress has been witnessed in image synthesis. Prior works mostly rely on pre-trained linguistic models, but a text is often too abstract to properly specify all the spatial properties of an image, e.g., the layout configuration of a scene, leading to the sub-optimal results of complex scene generation. In this paper, we achieve accurate complex scene generation by proposing a semantically controllable Layout-AWare diffusion model, termed LAW-Diffusion. Distinct from the previous Layout-to-Image generation (L2I) methods that primarily explore category-aware relationships, LAW-Diffusion introduces a spatial dependency parser to encode the location-aware semantic coherence across objects as a layout embedding and produces a scene with perceptually harmonious object styles and contextual relations. To be specific, we delicately instantiate each object's regional semantics as an object region map and leverage a location-aware cross-object attention module to capture the spatial dependencies among those disentangled representations. We further propose an adaptive guidance schedule for our layout guidance to mitigate the trade-off between the regional semantic alignment and the texture fidelity of generated objects. Moreover, LAW-Diffusion allows for instance reconfiguration while maintaining the other regions in a synthesized image by introducing a layout-aware latent grafting mechanism to recompose its local regional semantics. To better verify the plausibility of generated scenes, we propose a new evaluation metric for the L2I task, dubbed Scene Relation Score (SRS) to measure how the images preserve the rational and harmonious relations among contextual objects. Comprehensive experiments on COCO-Stuff and Visual-Genome demonstrate that our LAW-Diffusion yields the state-of-the-art generative performance, especially with coherent object relations.
****
Multi-Label Knowledge Distillation
Penghui Yang, Ming-Kun Xie, Chen-Chen Zong, Lei Feng, Gang Niu, Masashi Sugiyama, Sheng-Jun Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17271-17280
Existing knowledge distillation methods typically work by imparting the knowledge of output logits or intermediate feature maps from the teacher network to the student network, which is very successful in multi-class single-label learning. However, these methods can hardly be extended to the multi-label learning scenario, where each instance is associated with multiple semantic labels, because the prediction probabilities do not sum to one and feature maps of the whole example may ignore minor classes in such a scenario. In this paper, we propose a novel multi-label knowledge distillation method. On one hand, it exploits the informa

tive semantic knowledge from the logits by dividing the multi-label learning pro blem into a set of binary classification problems; on the other hand, it enhance s the distinctiveness of the learned feature representations by leveraging the s tructural information of label-wise embeddings. Experimental results on multiple benchmark datasets validate that the proposed method can avoid knowledge counte raction among labels, thus achieving superior performance against diverse compar ing methods.
****
Towards Geospatial Foundation Models via Continual Pretraining
Matías Mendieta, Boran Han, Xingjian Shi, Yi Zhu, Chen Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16806-168 16
Geospatial technologies are becoming increasingly essential in our world for a w ide range of applications, including agriculture, urban planning, and disaster r esponse. To help improve the applicability and performance of deep learning mode ls on these geospatial tasks, various works have begun investigating foundation models for this domain. Researchers have explored two prominent approaches for i ntroducing such models in geospatial applications, but both have drawbacks in te rms of limited performance benefit or prohibitive training cost. Therefore, in t his work, we propose a novel paradigm for building highly effective geospatial f oundation models with minimal resource cost and carbon impact. We first construc t a compact yet diverse dataset from multiple sources to promote feature diversi ty, which we term GeoPile. Then, we investigate the potential of continual pretr aining from large-scale ImageNet-22k models and propose a multi-objective contin ual pretraining paradigm, which leverages the strong representations of ImageNet while simultaneously providing the freedom to learn valuable in-domain features . Our approach outperforms previous state-of-the-art geospatial pretraining meth ods in an extensive evaluation on seven downstream datasets covering various tas ks such as change detection, classification, multi-label classification, semanti c segmentation, and super-resolution. Code is available at https://github.com/mm endiet/GFM.
****
ConSlide: Asynchronous Hierarchical Interaction Transformer with Breakup-Reorgan ize Rehearsal for Continual Whole Slide Image Analysis
Yanyan Huang, Weiqin Zhao, Shujun Wang, Yu Fu, Yuming Jiang, Lequan Yu; Proceedi ngs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp . 21349-21360
Whole slide image (WSI) analysis has become increasingly important in the medica l imaging community, enabling automated and objective diagnosis, prognosis, and therapeutic-response prediction. However, in clinical practice, the continuous p rogress of evolving WSI acquisition technology, the diversity of scanners, and d ifferent imaging protocols hamper the utility of WSI analysis models. In this pa per, we propose the FIRST continual learning framework for WSI analysis, named C onSlide, to tackle the challenges of enormous image size, utilization of hierarc hical structure, and catastrophic forgetting by progressive model updating on mu ltiple sequential datasets. Our framework contains three key components. The Hie rarchical Interaction Transformer (HIT) is proposed to model and utilize the hie rarchical structural knowledge of WSI. The BreakupReorganize (BuRo) rehearsal me thod is developed for WSI data replay with efficient region storing buffer and W SI reorganizing operation. The asynchronous updating mechanism is devised to enc ourage the network to learn generic and specific knowledge respectively during t he replay stage, based on a nested cross-scale similarity learning (CSSL) module . We evaluated the proposed ConSlide on four public WSI datasets from TCGA proje cts. It performs best over other state-of-the-art methods with a fair WSI-based continual learning setting and achieves a better trade-off of the overall perfor mance and forgetting on previous tasks.
****
RepQ-ViT: Scale Reparameterization for Post-Training Quantization of Vision Tran sformers
Zhikai Li, Junrui Xiao, Lianwei Yang, Qingyi Gu; Proceedings of the IEEE/CVF Int

ernational Conference on Computer Vision (ICCV), 2023, pp. 17227-17236
Post-training quantization (PTQ), which only requires a tiny dataset for calibration without end-to-end retraining, is a light and practical model compression technique. Recently, several PTQ schemes for vision transformers (ViTs) have been presented; unfortunately, they typically suffer from non-trivial accuracy degradation, especially in low-bit cases. In this paper, we propose RepQ-ViT, a novel PTQ framework for ViTs based on quantization scale reparameterization, to address the above issues. RepQ-ViT decouples the quantization and inference processes, where the former employs complex quantizers and the latter employs scale-reparameterized simplified quantizers. This ensures both accurate quantization and efficient inference, which distinguishes it from existing approaches that sacrifice quantization performance to meet the target hardware. More specifically, we focus on two components with extreme distributions: post-LayerNorm activations with severe inter-channel variation and post-Softmax activations with power-law features, and initially apply channel-wise quantization and log(sqrt(2)) quantization, respectively. Then, we reparameterize the scales to hardware-friendly layer-wise quantization and log2 quantization for inference, with only slight accuracy or computational costs. Extensive experiments are conducted on multiple vision tasks with different model variants, proving that RepQ-ViT, without hyperparameters and expensive reconstruction procedures, can outperform existing strong baselines and encouragingly improve the accuracy of 4-bit PTQ of ViTs to a usable level. Code is available at https://github.com/zkkli/RepQ-ViT.
****
Enhancing Privacy Preservation in Federated Learning via Learning Rate Perturbation
Guangnian Wan, Haitao Du, Xuejing Yuan, Jun Yang, Meiling Chen, Jie Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4772-4781
Federated learning (FL) is a privacy-enhanced distributed machine learning framework, in which multiple clients collaboratively train a global model by exchanging their model updates without sharing local private data. However, the adversary can use gradient inversion attacks to reveal the clients' privacy from the shared model updates. Previous attacks assume the adversary can infer the local learning rate of each client, while we observe that: (1) using the uniformly distributed random local learning rates does not incur much accuracy loss of the global model, and (2) personalizing local learning rates can mitigate the drift issue which is caused by non-IID (identically and independently distributed) data. Moreover, we theoretically derive a convergence guarantee to FedAvg with uniformly perturbed local learning rates. Therefore, by perturbing the learning rate of each client with random noise, we propose a learning rate perturbation (LRP) defense against gradient inversion attacks. Specifically, for classification tasks, we adapt LPR to ada-LPR by personalizing the expectation of each local learning rate. The experiments show that our defenses can well enhance privacy preservation against existing gradient inversion attacks, and LRP outperforms 5 baseline defenses against a state-of-the-art gradient inversion attack. In addition, our defenses only incur minor accuracy reductions (less than 0.5%) of the global model. So they are effective in real applications.
****
UMC: A Unified Bandwidth-efficient and Multi-resolution based Collaborative Perception Framework
Tianhang Wang, Guang Chen, Kai Chen, Zhengfa Liu, Bo Zhang, Alois Knoll, Changjun Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8187-8196
Multi-agent collaborative perception (MCP) has recently attracted much attention. It includes three key processes: communication for sharing, collaboration for integration, and reconstruction for different downstream tasks. Existing methods pursue designing the collaboration process alone, ignoring their intrinsic interactions and resulting in suboptimal performance. In contrast, we aim to propose a Unified Collaborative perception framework named UMC, optimizing the communication, collaboration, and reconstruction processes with the Multi-resolution tec

hnique. The communication introduces a novel trainable multi-resolution and sele
ctive-region (MRSR) mechanism, achieving higher quality and lower bandwidth. The
n, a graph-based collaboration is proposed, conducting on each resolution to ada
pt the MRSR. Finally, the reconstruction integrates the multi-resolution collabo
rative features for downstream tasks. Since the general metric can not reflect t
he performance enhancement brought by MCP systematically, we introduce a brand-n
ew evaluation metric that evaluates the MCP from different perspectives. To veri
fy our algorithm, we conducted experiments on the V2X-Sim and OPV2V datasets. Ou
r quantitative and qualitative experiments prove that the proposed UMC outperfor
ms the state-of-the-art collaborative perception approaches.
****

Viewing Graph Solvability in Practice
Federica Arrigoni, Tomas Pajdla, Andrea Fusiello; Proceedings of the IEEE/CVF In
ternational Conference on Computer Vision (ICCV), 2023, pp. 8147-8155
We present an advance in understanding the projective Structure-from-Motion, foc
using in particular on the viewing graph: such a graph has cameras as nodes and
fundamental matrices as edges. We propose a practical method for testing finite
solvability, i.e., whether a viewing graph induces a finite number of camera con
figurations. Our formulation uses a significantly smaller number of equations (u
p to 400x) with respect to previous work. As a result, this is the only method i
n the literature that can be applied to large viewing graphs coming from real da
tasets, comprising up to 300K edges. In addition, we develop the first algorithm
 for identifying maximal finite-solvable components.
****

SATR: Zero-Shot Semantic Segmentation of 3D Shapes
Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, Peter Wonka; Proceedings o
f the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 151
66-15179
We explore the task of zero-shot semantic segmentation of 3D shapes by using lar
ge-scale off-the-shelf 2D im- age recognition models. Surprisingly, we find that
 modern zero-shot 2D object detectors are better suited for this task than conte
mporary text/image similarity predictors or even zero-shot 2D segmentation netwo
rks. Our key finding is that it is possible to extract accurate 3D segmentation
maps from multi-view bounding box predictions by using the topological propertie
s of the underlying surface. For this, we develop the Segmentation Assignment wi
th Topological Reweighting (SATR) algorithm and evaluate it on ShapeNetPart and
our proposed FAUST benchmarks. SATR achieves state-of-the-art performance and ou
tperforms a baseline algorithm by 1.3% and 4% average mIoU on the FAUST coarse a
nd fine-grained benchmarks, respectively, and by 5.2% average mIoU on the ShapeN
etPart bench- mark. Our source code and data will be publicly released. Project
webpage: https://samir55.github.io/SATR/.
****

ReactioNet: Learning High-Order Facial Behavior from Universal Stimulus-Reaction
 by Dyadic Relation Reasoning
Xiaotian Li, Taoyue Wang, Geran Zhao, Xiang Zhang, Xi Kang, Lijun Yin; Proceedin
gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp.
 20774-20785
Diverse visual stimuli can evoke various human affective states, which are usual
ly manifested in an individual's muscular actions and facial expressions. In lab
-controlled emotion datasets, such a critical component (i.e., stimulus) was com
monly designed in a limited way, making researchers incapable of generalizing th
e universal correlation and causation of stimulus-reaction as well as predicting
 possible emotions from context, timing, and relation. In this paper, we collect
ed a large-scale spontaneous facial behavior database ReactioNet, which contains
 1.1 million coupled stimulus-reaction tuples (visual/audio/caption from both st
imuli and subjects). We introduce a new facial behavior detection scenario, Dyad
ic Relation Reasoning (DRR), which aims to detect facial actions by reasoning th
eir relations with stimuli. By aggregating the dyadic information, our method es
sentially forms a relation prototype Universal Stimulus Reaction (U-SR), which e
ncodes the low-order and high-order relationships between stimulus agents and fa

cial reactions. A framework with both non-graph and graph modules is further developed to evaluate DRR-based facial action unit detection, facial expression recognition, and scene classification. Specifically, to learn "what" arouses a facial reaction, the non-graph module associates and projects the fine-grained stimulus-reaction features into common subspaces using cross-domain contrastive learning. To learn "how" stimulus-reaction are mutually related, the graph module adopts Graph Convolution Network to represent, converge, and infer the dyadic U-SR relation under two relation assumptions (i.e., homophily and heterophily). Extensive experiments demonstrate the effectiveness of the proposed work. The new dataset will be available for the research community.
****
Pseudo Flow Consistency for Self-Supervised 6D Object Pose Estimation
Yang Hai, Rui Song, Jiaojiao Li, David Ferstl, Yinlin Hu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14075-14085
Most self-supervised 6D object pose estimation methods can only work with additional depth information or rely on the accurate annotation of 2D segmentation masks, limiting their application range. In this paper, we propose a 6D object pose estimation method that can be trained with pure RGB images without any auxiliary information. We first obtain a rough pose initialization from networks trained on synthetic images rendered from the target's 3D mesh. Then, we introduce a refinement strategy leveraging the geometry constraint in synthetic-to-real image pairs from multiple different views. We formulate this geometry constraint as pixel-level flow consistency between the training images with dynamically generated pseudo labels. We evaluate our method on three challenging datasets and demonstrate that it outperforms state-of-the-art self-supervised methods significantly, with neither 2D annotations nor additional depth images.
****
Emotional Listener Portrait: Neural Listener Head Generation with Emotion
Luchuan Song, Guojun Yin, Zhenchao Jin, Xiaoyi Dong, Chenliang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20839-20849
Listener head generation centers on generating non-verbal behaviors (e.g., smile) of a listener in reference to the information delivered by a speaker. A significant challenge when generating such responses is the non-deterministic nature of fine-grained facial expressions during a conversation, which varies depending on the emotions and attitudes of both the speaker and the listener. To tackle this problem, we propose the Emotional Listener Portrait (ELP), which treats each fine-grained facial motion as a composition of several discrete motion-codewords and explicitly models the probability distribution of the motions under different emotional contexts in conversation. Benefiting from the "explicit" and "discrete" design, our ELP model can not only automatically generate natural and diverse responses toward a given speaker via sampling from the learned distribution but also generate controllable responses with a predetermined attitude. Under several quantitative metrics, our ELP exhibits significant improvements compared to previous methods.
****
Unsupervised Domain Adaptation for Training Event-Based Networks Using Contrastive Learning and Uncorrelated Conditioning
Dayuan Jian, Mohammad Rostami; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18721-18731
Event-based cameras offer reliable measurements for preforming computer vision tasks in high-dynamic range environments and during fast motion maneuvers. However, adopting deep learning in event-based vision faces the challenge of annotated data scarcity due to recency of event cameras. Transferring the knowledge that can be obtained from conventional camera annotated data offers a practical solution to this challenge. We develop an unsupervised domain adaptation algorithm for training a deep network for event-based data image classification using contrastive learning and uncorrelated conditioning of data. Our solution outperforms the existing algorithms for this purpose.
****

Probabilistic Human Mesh Recovery in 3D Scenes from Egocentric Views
Siwei Zhang, Qianli Ma, Yan Zhang, Sadegh Aliakbarian, Darren Cosker, Siyu Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7989-8000
Automatic perception of human behaviors during social interactions is crucial for AR/VR applications, and an essential component is estimation of plausible 3D human pose and shape of our social partners from the egocentric view. One of the biggest challenges of this task is severe body truncation due to close social distances in egocentric scenarios, which brings large pose ambiguities for unseen body parts. To tackle this challenge, we propose a novel scene-conditioned diffusion method to model the body pose distribution. Conditioned on the 3D scene geometry, the diffusion model generates bodies in plausible human-scene interactions, with the sampling guided by a physics-based collision score to further resolve human-scene interpenetrations. The classifier-free training enables flexible sampling with different conditions and enhanced diversity. A visibility-aware graph convolution model guided by per-joint visibility serves as the diffusion denoiser to incorporate inter-joint dependencies and per-body-part control. Extensive evaluations show that our method generates bodies in plausible interactions with 3D scenes, achieving both superior accuracy for visible joints and diversity for invisible body parts. The code is available at https://sanweiliti.github.io/egohmr/egohmr.html.
****
ImGeoNet: Image-induced Geometry-aware Voxel Representation for Multi-view 3D Object Detection
Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, Min Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6996-7007
We propose ImGeoNet, a multi-view image-based 3D object detection framework that models a 3D space by an image-induced geometry-aware voxel representation.
Unlike previous methods which aggregate 2D features into 3D voxels without considering geometry, ImGeoNet learns to induce geometry from multi-view images to alleviate the confusion arising from voxels of free space, and during the inference phase, only images from multiple views are required.
Besides, a powerful pre-trained 2D feature extractor can be leveraged by our representation, leading to a more robust performance.
To evaluate the effectiveness of ImGeoNet, we conduct quantitative and qualitative experiments on three indoor datasets, namely ARKitScenes, ScanNetV2, and ScanNet200.
The results demonstrate that ImGeoNet outperforms the current state-of-the-art multi-view image-based method, ImVoxelNet, on all three datasets in terms of detection accuracy.
In addition, ImGeoNet shows great data efficiency by achieving results comparable to ImVoxelNet with 100 views while utilizing only 40 views.
Furthermore, our studies indicate that our proposed image-induced geometry-aware representation can enable image-based methods to attain superior detection accuracy than the seminal point cloud-based method, VoteNet, in two practical scenarios: (1) scenarios where point clouds are sparse and noisy, such as in ARKitScenes, and (2) scenarios involve diverse object classes, particularly classes of small objects, as in the case in ScanNet200.
****
DRAW: Defending Camera-shooted RAW Against Image Manipulation
Xiaoxiao Hu, Qichao Ying, Zhenxing Qian, Sheng Li, Xinpeng Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22434-22444
RAW files are the initial measurement of scene radiance widely used in most cameras, and the ubiquitously-used RGB images are converted from RAW data through Image Signal Processing (ISP) pipelines. Nowadays, digital images are risky of being nefariously manipulated. Inspired by the fact that innate immunity is the first line of body defense, we propose DRAW, a novel scheme of defending images against manipulation by protecting their sources, i.e., camera-shooted RAWs. Specif

ically, we design a lightweight Multi-frequency Partial Fusion Network (MPF-Net) friendly to devices with limited computing resources by frequency learning and partial feature fusion. It introduces invisible watermarks as protective signal into the RAW data. The protection capability can not only be transferred into the rendered RGB images regardless of the applied ISP pipeline, but also is resilient to post-processing operations such as blurring or compression. Once the image is manipulated, we can accurately identify the forged areas with a localization network. Extensive experiments on several famous RAW datasets, e.g., RAISE, FiveK and SIDD, indicate the effectiveness of our method. We hope that this technique can be used in future cameras as an option for image protection, which could effectively restrict image manipulation at the source.
****
Controllable Person Image Synthesis with Pose-Constrained Latent Diffusion
Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, Tao Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22768-22777
Controllable person image synthesis aims at rendering a source image based on user-specified changes in body pose or appearance. Prior art approaches leverage pixel-level denoising diffusion models conditioned on the coarse skeleton via cross-attention. This leads to two limitations: low efficiency and inaccurate condition information. To address both issues, a novel Pose-Constrained Latent Diffusion model (PoCoLD) is introduced. Rather than using the skeleton as a sparse pose representation, we exploit DensePose which offers much richer body structure information. To effectively capitalize DensePose at a low cost, we propose an efficient pose-constrained attention module that is capable of modeling the complex interplay between appearance and pose. Extensive experiments show that our PoCoLD outperforms the state-of-the-art competitors in image synthesis fidelity. Critically, it runs 2x faster and consumes 3.6x smaller memory than the latest diffusion-model-based alternative during inference.
****
Diffusion-based Image Translation with Label Guidance for Domain Adaptive Semantic Segmentation
Duo Peng, Ping Hu, Qiuhong Ke, Jun Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 808-820
Translating images from a source domain to a target domain for learning target models is one of the most common strategies in domain adaptive semantic segmentation (DASS). However, existing methods still struggle to preserve semantically-consistent local details between the original and translated images. In this work, we present an innovative approach that addresses this challenge by using source-domain labels as explicit guidance during image translation. Concretely, we formulate cross-domain image translation as a denoising diffusion process and utilize a novel Semantic Gradient Guidance (SGG) method to constrain the translation process, conditioning it on the pixel-wise source labels. Additionally, a Progressive Translation Learning (PTL) strategy is devised to enable the SGG method to work reliably across domains with large gaps. Extensive experiments demonstrate the superiority of our approach over state-of-the-art methods.
****
TopoSeg: Topology-Aware Nuclear Instance Segmentation
Hongliang He, Jun Wang, Pengxu Wei, Fan Xu, Xiangyang Ji, Chang Liu, Jie Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21307-21316
Nuclear instance segmentation has been critical for pathology image analysis in medical science, e.g., cancer diagnosis. Current methods typically adopt pixel-wise optimization for nuclei boundary exploration, where rich structural information could be lost for subsequent quantitative morphology assessment. To address this issue, we develop a topology-aware segmentation approach, termed TopoSeg, which exploits topological structure information to keep the predictions rational, especially in common situations with densely touching and overlapping nucleus instances. Concretely, TopoSeg builds on a topology-aware module (TAM), which encodes dynamic changes of different topology structures within the three-class pr

obability maps (inside, boundary, and background) of the nuclei to persistence barcodes and makes the topology-aware loss function. To efficiently focus on regions with high topological errors, we propose an adaptive topology-aware selection (ATS) strategy to enhance the topology-aware optimization procedure further. Experiments on three nuclear instance segmentation datasets justify the superiority of TopoSeg, which achieves state-of-the-art performance. The code is available at https://github.com/hhlisme/toposeg.
****
SceneRF: Self-Supervised Monocular 3D Scene Reconstruction with Radiance Fields
Anh-Quan Cao, Raoul de Charette; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9387-9398
3D reconstruction from a single 2D image was extensively covered in the literature but relies on depth supervision at training time, which limits its applicability. To relax the dependence to depth we propose SceneRF, a self-supervised monocular scene reconstruction method using only posed image sequences for training. Fueled by the recent progress in neural radiance fields (NeRF) we optimize a radiance field though with explicit depth optimization and a novel probabilistic sampling strategy to efficiently handle large scenes. At inference, a single input image suffices to hallucinate novel depth views which are fused together to obtain 3D scene reconstruction. Thorough experiments demonstrate that we outperform all baselines for novel depth views synthesis and scene reconstruction, on indoor BundleFusion and outdoor SemanticKITTI. Code is available at https://astra-vision.github.io/SceneRF .
****
Isomer: Isomerous Transformer for Zero-shot Video Object Segmentation
Yichen Yuan, Yifan Wang, Lijun Wang, Xiaoqi Zhao, Huchuan Lu, Yu Wang, Weibo Su, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 966-976
Recent leading zero-shot video object segmentation (ZVOS) works devote to integrating appearance and motion information by elaborately designing feature fusion modules and identically applying them in multiple feature stages. Our preliminary experiments show that with the strong long-range dependency modeling capacity of Transformer, simply concatenating the two modality features and feeding them to vanilla Transformers for feature fusion can distinctly benefit the performance but at a cost of heavy computation. Through further empirical analysis, we find that attention dependencies learned in Transformer in different stages exhibit completely different properties: global query-independent dependency in the low-level stages and semantic-specific dependency in the high-level stages. Motivated by the observations, we propose two Transformer variants: i) Context-Sharing Transformer (CST) that learns the global-shared contextual information within image frames with a lightweight computation. ii) Semantic Gathering-Scattering Transformer (SGST) that models the semantic correlation separately for the foreground and background and reduces the computation cost with a soft token merging mechanism. We apply CST and SGST for low-level and high-level feature fusions, respectively, formulating a level-isomerous Transformer framework for ZVOS task. Compared with the baseline that uses vanilla Transformers for multi-stage fusion, ours significantly increase the speed by 13 times and achieves new state-of-the-art ZVOS performance. Code is available at https://github.com/DLUT-yyc/Isomer.
****
CPCM: Contextual Point Cloud Modeling for Weakly-supervised Point Cloud Semantic Segmentation
Lizhao Liu, Zhuangwei Zhuang, Shangxin Huang, Xunlong Xiao, Tianhang Xiang, Cen Chen, Jingdong Wang, Mingkui Tan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18413-18422
We study the task of weakly-supervised point cloud semantic segmentation with sparse annotations (e.g., less than 0.1% points are labeled), aiming to reduce the expensive cost of dense annotations. Unfortunately, with extremely sparse annotated points, it is very difficult to extract both contextual and object information for scene understanding such as semantic segmentation. Motivated by masked modeling (e.g., MAE) in image and video representation learning, we seek to endow

the power of masked modeling to learn contextual information from sparsely-annotated points. However, directly applying MAE to 3D point clouds with sparse annotations may fail to work. First, it is nontrivial to effectively mask out the informative visual context from 3D point clouds. Second, how to fully exploit the sparse annotations for context modeling remains an open question. In this paper, we propose a simple yet effective Contextual Point Cloud Modeling (CPCM) method that consists of two parts: a region-wise masking (RegionMask) strategy and a contextual masked training (CMT) method. Specifically, RegionMask masks the point cloud continuously in geometric space to construct a meaningful masked prediction task for subsequent context learning. CMT disentangles the learning of supervised segmentation and unsupervised masked context prediction for effectively learning the very limited labeled points and mass unlabeled points, respectively. Extensive experiments on the widely-tested ScanNet V2 and S3DIS benchmarks demonstrate the superiority of CPCM over the state-of-the-art.
****

PATMAT: Person Aware Tuning of Mask-Aware Transformer for Face Inpainting

Saman Motamed, Jianjin Xu, Chen Henry Wu, Christian Häne, Jean-Charles Bazin, Fernando De la Torre; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22778-22787

Generative models such as StyleGAN2 and Stable Diffusion have achieved state-of-the-art performance in computer vision tasks such as image synthesis, inpainting, and de-noising. However, current generative models for face inpainting often fail to preserve fine facial details and the identity of the person, despite creating aesthetically convincing image structures and textures. In this work, we propose Person Aware Tuning (PAT) of Mask-Aware Transformer (MAT) for face inpainting, which addresses this issue. Our proposed method, PATMAT, effectively preserves identity by incorporating reference images of a subject and fine-tuning a MAT architecture trained on faces. By using 40 reference images, PATMAT creates anchor points in MAT's style module, and tunes the model using the fixed anchors to adapt the model to a new face identity. Moreover, PATMAT's use of multiple images per anchor during training allows the model to use fewer reference images than competing methods. We demonstrate that PATMAT outperforms state-of-the-art models in terms of image quality, the preservation of person-specific details, and the identity of the subject. Our results suggest that PATMAT can be a promising approach for improving the quality of personalized face inpainting.
****

Adaptive Nonlinear Latent Transformation for Conditional Face Editing

Zhizhong Huang, Siteng Ma, Junping Zhang, Hongming Shan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21022-21031

Recent works for face editing usually manipulate the latent space of StyleGAN via the linear semantic directions. However, they usually suffer from the entanglement of facial attributes, need to tune the optimal editing strength, and are limited to binary attributes with strong supervision signals. This paper proposes a novel adaptive nonlinear latent transformation for disentangled and conditional face editing, termed AdaTrans. Specifically, our AdaTrans divides the manipulation process into several finer steps; i.e., the direction and size at each step are conditioned on both the facial attributes and the latent codes. In this way, AdaTrans describes an adaptive nonlinear transformation trajectory to manipulate the faces into target attributes while keeping other attributes unchanged. Then, AdaTrans leverages a predefined density model to constrain the learned trajectory in the distribution of latent codes by maximizing the likelihood of transformed latent code. Moreover, we also propose a disentangled learning strategy under a mutual information framework to eliminate the entanglement among attributes, which can further relax the need for labeled data. Consequently, AdaTrans enables a controllable face editing with the advantages of disentanglement, flexibility with non-binary attributes, and high fidelity. Extensive experimental results on various facial attributes demonstrate the qualitative and quantitative effectiveness of the proposed AdaTrans over existing state-of-the-art methods, especially in the most challenging scenarios with a large age gap and few labeled examples. The source code is available at https://github.com/Hzzone/AdaTrans.

****
Tiny Updater: Towards Efficient Neural Network-Driven Software Updating
Linfeng Zhang, Kaisheng Ma; Proceedings of the IEEE/CVF International Conference
 on Computer Vision (ICCV), 2023, pp. 23447-23459
Significant advancements have been accomplished with deep neural networks in div
erse visual tasks, which have substantially elevated their deployment in edge de
vice software. However, during the update of neural network-based software, user
s are required to download all the parameters of the neural network anew, which
harms the user experience. Motivated by previous progress in model compression,
we propose a novel training methodology named Tiny Updater to address this issue
. Specifically, by adopting the variant of pruning and knowledge distillation me
thods, Tiny Updater can update the neural network-based software by only downloa
ding a few parameters (10% 20%) instead of all the parameters in the neural netw
ork. Experiments on eleven datasets of three tasks, including image classificati
on, image-to-image translation, and video recognition have demonstrated its effe
ctiveness. Codes have been released in https://github.com/ArchipLab-LinfengZhang
/TinyUpdater.
****
INT2: Interactive Trajectory Prediction at Intersections
Zhijie Yan, Pengfei Li, Zheng Fu, Shaocong Xu, Yongliang Shi, Xiaoxue Chen, Yuha
ng Zheng, Yang Li, Tianyu Liu, Chuxuan Li, Nairui Luo, Xu Gao, Yilun Chen, Zuoxu
 Wang, Yifeng Shi, Pengfei Huang, Zhengxiao Han, Jirui Yuan, Jiangtao Gong, Guyu
e Zhou, Hang Zhao, Hao Zhao; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2023, pp. 8536-8547
Motion forecasting is an important component in autonomous driving systems. One
of the most challenging problems in motion forecasting is interactive trajectory
 prediction, whose goal is to jointly forecasts the future trajectories of inter
acting agents. To this end, we present a large-scale interactive trajectory pred
iction dataset named INT2 for INTeractive trajectory prediction at INTersections
. INT2 includes 612,000 scenes, each lasting 1 minute, containing up to 10,200 h
ours of data. The agent trajectories are auto-labeled by a high-performance offl
ine temporal detection and fusion algorithm, whose quality is further inspected
by human judges. Vectorized semantic maps and traffic light information are also
 included in INT2. Additionally, the dataset poses an interesting domain mismatc
h challenge.
 For each intersection, we treat rush-hour and non-rush-hour segments as differe
nt domains. We benchmark the best open-sourced interactive trajectory prediction
 method on INT2 and Waymo Open Motion, under in-domain and cross-domain settings
. The dataset, code and models are publicly available at https://github.com/AIR-
DISCOVER/INT2.
****
MapPrior: Bird's-Eye View Map Layout Estimation with Generative Models
Xiyue Zhu, Vlas Zyrianov, Zhijian Liu, Shenlong Wang; Proceedings of the IEEE/CV
F International Conference on Computer Vision (ICCV), 2023, pp. 8228-8239
Despite tremendous advancements in bird's-eye view (BEV) perception, existing mo
dels fall short in generating realistic and coherent semantic map layouts, and t
hey fail to account for uncertainties arising from partial sensor information (s
uch as occlusion or limited coverage). In this work, we introduce MapPrior, a no
vel BEV perception framework that combines a traditional discriminative BEV perc
eption model with a learned generative model for semantic map layouts. Our MapPr
ior delivers predictions with better accuracy, realism, and uncertainty awarenes
s. We evaluate our model on the large-scale nuScenes benchmark. At the time of s
ubmission, MapPrior outperforms the strongest competing method, with significant
ly improved MMD and ECE scores in camera- and LiDAR-based BEV perception. Furthe
rmore, our method can be used to perpetually generate layouts with unconditional
 sampling.
****
CAD-Estate: Large-scale CAD Model Annotation in RGB Videos
Kevis-Kokitsi Maninis, Stefan Popov, Matthias Nießner, Vittorio Ferrari; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, p

p. 20189-20199
We propose a method for annotating videos of complex multi-object scenes with a globally-consistent 3D representation of the objects. We annotate each object with a CAD model from a database, and place it in the 3D coordinate frame of the scene with a 9-DoF pose transformation. Our method is semi-automatic and works on commonly-available RGB videos, without requiring a depth sensor. Many steps are performed automatically, and the tasks performed by humans are simple, well-specified, and require only limited reasoning in 3D. This makes them feasible for crowd-sourcing and has allowed us to construct a large-scale dataset by annotating real-estate videos from YouTube. Our dataset CAD-Estate offers 101k instances of 12k unique CAD models placed in the 3D representations of 20k videos. In comparison to Scan2CAD, the largest existing dataset with CAD model annotations on real
scenes, CAD-Estate has 7x more instances and 4x more unique CAD models. We showcase the benefits of pre-training a Mask2CAD model on CAD-Estate for the task of automatic
3D object reconstruction and pose estimation, demonstrating that it leads to performance improvements on the popular Scan2CAD benchmark. The dataset is available at https://github.com/google-research/cad-estate.
****
Conditional Cross Attention Network for Multi-Space Embedding without Entanglement in Only a SINGLE Network
Chull Hwan Song, Taebaek Hwang, Jooyoung Yoon, Shunghyun Choi, Yeong Hyeon Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11112-11121
Many studies in vision tasks have aimed to create effective embedding spaces for single-label object prediction within an image. However, in reality, most objects possess multiple specific attributes, such as shape, color, and length, with each attribute composed of various classes. To apply models in real-world scenarios, it is essential to be able to distinguish between the granular components of an object. Conventional approaches to embedding multiple specific attributes into a single network often result in entanglement, where fine-grained features of each attribute cannot be identified separately.
To address this problem, we propose a Conditional Cross-Attention Network that induces disentangled multi-space embeddings for various specific attributes with only a single backbone. Firstly, we employ a cross-attention mechanism to fuse and switch the information of conditions (specific attributes), and we demonstrate its effectiveness through a diverse visualization example. Secondly, we leverage the vision transformer for the first time to a fine-grained image retrieval task and present a simple yet effective framework compared to existing methods. Unlike previous studies where performance varied depending on the benchmark dataset, our proposed method achieved consistent state-of-the-art performance on the FashionAI, DARN, DeepFashion, and Zappos50K benchmark datasets.
****
MB-TaylorFormer: Multi-Branch Efficient Transformer Expanded by Taylor Formula for Image Dehazing
Yuwei Qiu, Kaihao Zhang, Chenxi Wang, Wenhan Luo, Hongdong Li, Zhi Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12802-12813
In recent years, Transformer networks are beginning to replace pure convolutional neural networks (CNNs) in the field of computer vision due to their global receptive field and adaptability to input. However, the quadratic computational complexity of softmax-attention limits the wide application in image dehazing task, especially for high-resolution images. To address this issue, we propose a new Transformer variant, which applies the Taylor expansion to approximate the softmax-attention and achieves linear computational complexity. A multi-scale attention refinement module is proposed as a complement to correct the error of the Taylor expansion. Furthermore, we introduce a multi-branch architecture with multi-scale patch embedding to the proposed Transformer, which embeds features by overlapping deformable convolution of different scales. The design of multi-scale pa

tch embedding is based on three key ideas: 1) various sizes of the receptive field; 2) flexible shapes of the receptive field; 3) multi-level semantic information. Our model, named Multi-branch Transformer expanded by Taylor formula (MB-TaylorFormer), can embed coarse to fine features more flexibly at the patch embedding stage and capture long-distance pixel interactions with limited computational cost. Experimental results on several dehazing benchmarks show that MB-TaylorFormer achieves state-of-the-art performance with a light computational burden.
****
X-Mesh: Towards Fast and Accurate Text-driven 3D Stylization via Dynamic Textual Guidance
Yiwei Ma, Xiaoqing Zhang, Xiaoshuai Sun, Jiayi Ji, Haowei Wang, Guannan Jiang, Weilin Zhuang, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2749-2760
Text-driven 3D stylization is a complex and crucial task in the fields of computer vision (CV) and computer graphics (CG), aimed at transforming a bare mesh to fit a target text. Prior methods adopt text-independent multilayer perceptrons (MLPs) to predict the attributes of the target mesh with the supervision of CLIP loss. However, such text-independent architecture lacks textual guidance during predicting attributes, thus leading to unsatisfactory stylization and slow convergence. To address these limitations, we present X-Mesh, an innovative text-driven 3D stylization framework that incorporates a novel Text-guided Dynamic Attention Module (TDAM). The TDAM dynamically integrates the guidance of the target text by utilizing text-relevant spatial and channel-wise attentions during vertex feature extraction, resulting in more accurate attribute prediction and faster convergence speed. Furthermore, existing works lack standard benchmarks and automated metrics for evaluation, often relying on subjective and non-reproducible user studies to assess the quality of stylized 3D assets. To overcome this limitation, we introduce a new standard text-mesh benchmark, namely MIT-30, and two automated metrics, which will enable future research to achieve fair and objective comparisons. Our extensive qualitative and quantitative experiments demonstrate that X-Mesh outperforms previous state-of-the-art methods.
****
Muscles in Action
Mia Chiquier, Carl Vondrick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22091-22101
Human motion is created by, and constrained by, our muscles. We take a first step at building computer vision methods that represent the internal muscle activity that causes motion. We present a new dataset, Muscles in Action (MIA), to to learn to incorporate muscle activity into human motion representations. The dataset consists of 12.5 hours of synchronized video and surface electromyography (sEMG) data of 10 subjects performing various exercises. Using this dataset, we learn a bidirectional representation that predicts muscle activation from video, and conversely, reconstructs motion from muscle activation. We evaluate our model on in-distribution subjects and exercises, as well as on out-of-distribution subjects and exercises. We demonstrate how advances in modeling both modalities jointly can serve as conditioning for muscularly consistent motion generation. Putting muscles into computer vision systems will enable richer models of virtual humans, with applications in sports, fitness, and AR/VR.
****
Large-Scale Person Detection and Localization Using Overhead Fisheye Cameras
Lu Yang, Liulei Li, Xueshi Xin, Yifan Sun, Qing Song, Wenguan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19961-19971
Location determination finds wide applications in daily life. Instead of existing efforts devoted to localizing tourist photos captured by perspective cameras, in this article, we focus on developing person positioning solutions using overhead fisheye cameras. Such solutions are advantageous in large field of view (FOV), low cost, anti-occlusion, and unaggressive work mode (without the necessity of cameras carried by persons). However, related studies are quite scarce, due to the paucity of data. To stimulate research in this exciting area, we present LO

AF, the first large-scale overhead fisheye dataset for person detection and loca lization. LOAF is built with many essential features, e.g., i) the data cover ab undant diversities in scenes, human pose, density, and location; ii) it contains currently the largest number of annotated pedestrian, i.e., 600K bounding boxes with ground-truth location information; iii) the body-boxes are labeled as radi us-aligned so as to fully address the positioning challenge. To approach localiz ation, we build a fisheye person detection network, which exploits the fisheye d istortions by a clever position embedding strategy and is trained to predict rad ius-aligned human boxes end-to-end. Then, the actual locations of the detected p ersons are calculated by a numerical solution on the fisheye model and camera al titude data. Extensive experiments on LOAF validate the superiority of our fishe ye detector w.r.t. previous methods, and show that our whole fisheye positioning solution is able to locate all persons in FOV with an accuracy of 0.5m, within 0.1s. Our dataset and code shall be released.
****
ViLTA: Enhancing Vision-Language Pre-training through Textual Augmentation
Weihan Wang, Zhen Yang, Bin Xu, Juanzi Li, Yankui Sun; Proceedings of the IEEE/C VF International Conference on Computer Vision (ICCV), 2023, pp. 3158-3169
Vision-language pre-training (VLP) methods are blossoming recently, and its cruc ial goal is to jointly learn visual and textual features via a transformer-based architecture, demonstrating promising improvements on a variety of vision-langu age tasks. Prior arts usually focus on how to align visual and textual features, but strategies for improving the robustness of model and speeding up model conv ergence are left insufficiently explored.

In this paper, we propose a novel method ViLTA, comprising of two components to further facilitate the model to learn fine-grained representations among image-text pairs. For Masked Language Modeling (MLM), we propose a cross-distillation method to generate soft labels to enhance the robustness of model, which allevia tes the problem of treating synonyms of masked words as negative samples in one-hot labels. For Image-Text Matching (ITM), we leverage the current language enco der to synthesize hard negatives based on the context of language input, encoura ging the model to learn high-quality representations by increasing the difficult y of the ITM task. By leveraging the above techniques, our ViLTA can achieve bet ter performance on various vision-language tasks. Extensive experiments on bench mark datasets demonstrate that the effectiveness of ViLTA and its promising pote ntial for vision-language pre-training.
****
All-to-Key Attention for Arbitrary Style Transfer
Mingrui Zhu, Xiao He, Nannan Wang, Xiaoyu Wang, Xinbo Gao; Proceedings of the IE EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23109-23119
Attention-based arbitrary style transfer studies have shown promising performanc e in synthesizing vivid local style details. They typically use the all-to-all a ttention mechanism---each position of content features is fully matched to all p ositions of style features. However, all-to-all attention tends to generate dist orted style patterns and has quadratic complexity, limiting the effectiveness an d efficiency of arbitrary style transfer. In this paper, we propose a novel all-to-key attention mechanism---each position of content features is matched to sta ble key positions of style features---that is more in line with the characterist ics of style transfer. Specifically, it integrates two newly proposed attention forms: distributed and progressive attention. Distributed attention assigns atte ntion to key style representations that depict the style distribution of local r egions; Progressive attention pays attention from coarse-grained regions to fine -grained key positions. The resultant module, dubbed StyA2K, shows extraordinary performance in preserving the semantic structure and rendering consistent style patterns. Qualitative and quantitative comparisons with state-of-the-art method s demonstrate the superior performance of our approach. Codes and models are ava ilable on https://github.com/LearningHx/StyA2K.
****
Learning to Distill Global Representation for Sparse-View CT

Zilong Li, Chenglong Ma, Jie Chen, Junping Zhang, Hongming Shan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21196-21207

Sparse-view computed tomography (CT)---using a small number of projections for tomographic reconstruction---enables much lower radiation dose to patients and accelerated data acquisition. The reconstructed images, however, suffer from strong artifacts, greatly limiting their diagnostic value. Current trends for sparse-view CT turn to the raw data for better information recovery. The resultant dual-domain methods, nonetheless, suffer from secondary artifacts, especially in ultra-sparse view scenarios, and their generalization to other scanners/protocols is greatly limited. A crucial question arises: have the image post-processing methods reached the limit? Our answer is not yet. In this paper, we stick to image post-processing methods due to great flexibility and propose global representation (GloRe) distillation framework for sparse-view CT, termed GloReDi. First, we propose to learn GloRe with Fourier convolution, so each element in GloRe has an image-wide receptive field. Second, unlike methods that only use the full-view images for supervision, we propose to distill GloRe from intermediate-view reconstructed images that are readily available but not explored in previous literature. The success of GloRe distillation is attributed to two key components: representation directional distillation to align the GloRe directions, and band-pass-specific contrastive distillation to gain clinically important details. Extensive experiments demonstrate the superiority of the proposed GloReDi over the state-of-the-art methods, including dual-domain ones. The source code is available at https://github.com/longzilicart/GloReDi.

****

FocalFormer3D: Focusing on Hard Instance for 3D Object Detection

Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, Jose M. Alvarez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8394-8405

False negatives (FN) in 3D object detection, e.g., missing predictions of pedestrians, vehicles, or other obstacles, can lead to potentially dangerous situations in autonomous driving. While being fatal, this issue is understudied in many current 3D detection methods. In this work, we propose Hard Instance Probing (HIP), a general pipeline that identifies FN in a multi-stage manner and guides the models to focus on excavating difficult instances. For 3D object detection, we instantiate this method as FocalFormer3D, a simple yet effective detector that excels at excavating difficult objects and improving prediction recall. FocalFormer3D features a multi-stage query generation to discover hard objects and a box-level transformer decoder to efficiently distinguish objects from massive object candidates. Experimental results on the nuScenes and Waymo datasets validate the superior performance of FocalFormer3D. The advantage leads to strong performance on both detection and tracking, in both LiDAR and multi-modal settings. Notably, FocalFormer3D achieves a 70.5 mAP and 73.9 NDS on nuScenes detection benchmark, while the nuScenes tracking benchmark shows 72.1 AMOTA, both ranking 1st place on the nuScenes LiDAR leaderboard. Our code is available at https://github.com/NVlabs/FocalFormer3D.

****

Not Every Side Is Equal: Localization Uncertainty Estimation for Semi-Supervised 3D Object Detection

Chuxin Wang, Wenfei Yang, Tianzhu Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3814-3824

Semi-supervised 3D object detection from point cloud aims to train a detector with a small number of labeled data and a large number of unlabeled data. The core of existing methods lies in how to select high-quality pseudo-labels using the designed quality evaluation criterion. However, these methods treat each pseudo bounding box as a whole and assign equal importance to each side during training, which is detrimental to model performance due to many sides having poor localization quality. Besides, existing methods filter out a large number of low-quality pseudo-labels, which also contain some correct regression values that can help with model training. To address the above issues, we propose a side-aware fram

ework for semi-supervised 3D object detection consisting of three key designs: a 3D bounding box parameterization method, an uncertainty estimation module, and a pseudo-label selection strategy. These modules work together to explicitly estimate the localization quality of each side and assign different levels of importance during the training phase. Extensive experiment results demonstrate that the proposed method can consistently outperform baseline models under different scenes and evaluation criteria. Moreover, our method achieves state-of-the-art performance on three datasets with different labeled ratios.
****
Teaching CLIP to Count to Ten
Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, Tali Dekel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3170-3180
Large vision-language models, such as CLIP, learn robust representations of text and images, facilitating advances in many downstream tasks, including zero-shot classification and text-to-image generation. However, these models have several well-documented limitations. They fail to encapsulate compositional concepts, such as counting. To the best of our knowledge, this work is the first to extend CLIP to handle object counting. We introduce a simple yet effective method to improve the quantitative understanding of vision-language models, while maintaining their overall performance on common benchmarks.
Our method automatically augments image captions to create hard negative samples that differ from the original captions by only the number of objects. For example, an image of three dogs can be contrasted with the negative caption "Six dogs playing in the yard". A dedicated loss encourages discrimination between the correct caption and its negative variant.
In addition, we introduce CountBench, a new benchmark for evaluating a model's understanding of object counting, and demonstrate significant improvement over baseline models on this task. Furthermore, we leverage our improved CLIP representations for text-conditioned image generation, and show that our model can produce specific counts of objects more reliably than existing ones.
****
TEMPO: Efficient Multi-View Pose Estimation, Tracking, and Forecasting
Rohan Choudhury, Kris M. Kitani, László A. Jeni; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14750-14760
Existing volumetric methods for predicting 3D human pose estimation are accurate, but computationally expensive and optimized for single time-step prediction. We present TEMPO, an efficient multi-view pose estimation model that learns a robust spatiotemporal representation, improving pose accuracy while also tracking and forecasting human pose. We significantly reduce computation compared to the state-of-the-art by recurrently computing per-person 2D pose features, fusing both spatial and temporal information into a single representation. In doing so, our model is able to use spatiotemporal context to predict more accurate human poses without sacrificing efficiency. We further use this representation to track human poses over time as well as predict future poses. Finally, we demonstrate that our model is able to generalize across datasets without scene-specific fine-tuning. TEMPO achieves 10% better MPJPE with a 33x improvement in FPS compared to TesseTrack on the challenging CMU Panoptic Studio dataset. Our code and demos are available at https://rccchoudhury.github.io/tempo2023.
****
SparseMAE: Sparse Training Meets Masked Autoencoders
Aojun Zhou, Yang Li, Zipeng Qin, Jianbo Liu, Junting Pan, Renrui Zhang, Rui Zhao, Peng Gao, Hongsheng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16176-16186
Masked Autoencoders (MAE) and its variants have proven to be effective for pretraining large-scale Vision Transformers (ViTs). However, small-scale models do not benefit from the pretraining mechanisms due to limited capacity. Sparse training is a method of transferring representations from large models to small ones by pruning unimportant parameters. However, naively combining MAE finetuning with sparse training make the network task-specific, resulting in the loss of task-a

gnostic knowledge, which is crucial for model generalization. In this paper, we
aim to reduce model complexity from large vision transformers pretrained by MAE
with assistant of sparse training. We summarize various sparse training methods
to prune large vision transformers during MAE pretraining and finetuning stages,
and discuss their shortcomings. To improve learning both task-agnostic and task
-specific knowledge, we propose SparseMAE, a novel two-stage sparse training met
hod that includes sparse pretraining and sparse finetuning. In sparse pretrainin
g, we dynamically prune a small-scale sub-network from a ViT-Base. During finetu
ning, the sparse sub-network adaptively changes its topology connections under t
he task-agnostic knowledge of the full model. Extensive experimental results dem
onstrate the effectiveness of our method and its superiority on small-scale visi
on transformers. Code will be available at https://github.com/aojunzz/SparseMAE.
****
DiffPose: SpatioTemporal Diffusion Model for Video-Based Human Pose Estimation
Runyang Feng, Yixing Gao, Tze Ho Elden Tse, Xueqing Ma, Hyung Jin Chang; Proceed
ings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, p
p. 14861-14872
Denoising diffusion probabilistic models that were initially proposed for realis
tic image generation have recently shown success in various perception tasks (e.
g., object detection and image segmentation) and are increasingly gaining attent
ion in computer vision. However, extending such models to multi-frame human pose
estimation is non-trivial due to the presence of the additional temporal dimens
ion in videos. More importantly, learning representations that focus on keypoint
regions is crucial for accurate localization of human joints. Nevertheless, the
adaptation of the diffusion-based methods remains unclear on how to achieve suc
h objective. In this paper, we present DiffPose, a novel diffusion architecture
that formulates video-based human pose estimation as a conditional heatmap gener
ation problem. First, to better leverage temporal information, we propose Spatio
Temporal Representation Learner which aggregates visual evidences across frames
and uses the resulting features in each denoising step as a condition. In additi
on, we present a mechanism called Lookup-based MultiScale Feature Interaction th
at determines the correlations between local joints and global contexts across m
ultiple scales. This mechanism generates delicate representations that focus on
keypoint regions. Altogether, by extending diffusion models, we show two unique
characteristics from DiffPose on pose estimation task: (i) the ability to combin
e multiple sets of pose estimates to improve prediction accuracy, particularly f
or challenging joints, and (ii) the ability to adjust the number of iterative st
eps for feature refinement without retraining the model. DiffPose sets new state
-of-the-art results on three benchmarks: PoseTrack2017, PoseTrack2018, and PoseT
rack21.
****
ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-I
mage Generation
Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, Wangmeng Zuo; Proce
edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 15943-15953
In addition to the unprecedented ability in imaginary creation, large text-to-im
age models are expected to take customized concepts in image generation. Existin
g works generally learn such concepts in an optimization-based manner, yet bring
ing excessive computation or memory burden. In this paper, we instead propose a
learning-based encoder, which consists of a global and a local mapping networks
for fast and accurate customized text-to-image generation. In specific, the glob
al mapping network projects the hierarchical features of a given image into mult
iple "new" words in the textual word embedding space, i.e., one primary word for
well-editable concept and other auxiliary words to exclude irrelevant disturban
ces (e.g., background). In the meantime, a local mapping network injects the enc
oded patch features into cross attention layers to provide omitted details, with
out sacrificing the editability of primary concepts. We compare our method with
existing optimization-based approaches on a variety of user-defined concepts, an
d demonstrate that our method enables highfidelity inversion and more robust edi

tability with a significantly faster encoding process. Our code is publicly avai
lable at https://github.com/csyxwei/ELITE.
****
Text2Performer: Text-Driven Human Video Generation
Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22747-22757
Text-driven content creation has evolved to be a transformative technique that revolutionizes creativity. Here we study the task of text-driven human video generation, where a video sequence is synthesized from texts describing the appearance and motions of a target performer. Compared to general text-driven video generation, human-centric video generation requires maintaining the appearance of synthesized human while performing complex motions. In this work, we present Text2Performer to generate vivid human videos with articulated motions from texts. Text2Performer has two novel designs: 1) decomposed human representation and 2) diffusion-based motion sampler. First, we decompose the VQVAE latent space into human appearance and pose representation in an unsupervised manner by utilizing the nature of human videos. In this way, the appearance is well maintained along the generated frames. Then, we propose continuous VQ-diffuser to sample a sequence of pose embeddings. Unlike existing VQ-based methods that operate in the discrete space, continuous VQ-diffuser directly outputs the continuous pose embeddings for better motion modeling. Finally, motion-aware masking strategy is designed to mask the pose embeddings spatial-temporally to enhance the temporal coherence. Moreover, to facilitate the task of text-driven human video generation, we contribute a Fashion-Text2Video dataset with manually annotated action labels and text descriptions. Extensive experiments demonstrate that Text2Performer generates high-quality human videos (up to 512x256 resolution) with diverse appearances and flexible motions. Our project page is https://yumingj.github.io/projects/Text2Performer.html
****
A Simple Recipe to Meta-Learn Forward and Backward Transfer
Edoardo Cetin, Antonio Carta, Oya Celiktutan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18732-18742
Meta-learning holds the potential to provide a general and explicit solution to tackle interference and forgetting in continual learning. However, many popular algorithms introduce expensive and unstable optimization processes with new key hyper-parameters and requirements, hindering their applicability. We propose a new, general, and simple meta-learning algorithm for continual learning (SiM4C) that explicitly optimizes to minimize forgetting and facilitate forward transfer. We show our method is stable, introduces only minimal computational overhead, and can be integrated with any memory-based continual learning algorithm in only a few lines of code. SiM4C meta-learns how to effectively continually learn even on very long task sequences, largely outperforming prior meta-approaches. Naively integrating with existing memory-based algorithms, we also record universal performance benefits and state-of-the-art results across different visual classification benchmarks without introducing new hyper-parameters.
****
4D Myocardium Reconstruction with Decoupled Motion and Shape Model
Xiaohan Yuan, Cong Liu, Yangang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21252-21262
Estimating the shape and motion state of the myocardium is essential in diagnosing cardiovascular diseases. However, cine magnetic resonance (CMR) imaging is dominated by 2D slices, whose large slice spacing challenges inter-slice shape reconstruction and motion acquisition. To address this problem, we propose a 4D reconstruction method that decouples motion and shape, which can predict the inter-/intra- shape and motion estimation from a given sparse point cloud sequence obtained from limited slices. Our framework comprises a neural motion model and an end-diastolic (ED) shape model. The implicit ED shape model can learn a continuous boundary and encourage the motion model to predict without the supervision of ground truth deformation, and the motion model enables canonical input of the s

hape model by deforming any point from any phase to the ED phase. Additionally, the constructed ED-space enables pre-training of the shape model, thereby guiding the motion model and addressing the issue of data scarcity. We propose the first 4D myocardial dataset as we know and verify our method on the proposed, public, and cross-modal datasets, showing superior reconstruction performance and enabling various clinical applications.
****
IntentQA: Context-aware Video Intent Reasoning
Jiapeng Li, Ping Wei, Wenjuan Han, Lifeng Fan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11963-11974
In this paper, we propose a novel task IntentQA, a special VideoQA task focusing on video intent reasoning, which has become increasingly important for AI with its advantages in equipping AI agents with the capability of reasoning beyond mere recognition in daily tasks. We also contribute a large-scale VideoQA dataset for this task. We propose a Context-aware Video Intent Reasoning model (CaVIR) consisting of i) Video Query Language (VQL) for better cross-modal representation of the situational context, ii) Contrastive Learning module for utilizing the contrastive context, and iii) Commonsense Reasoning module for incorporating the commonsense context. Comprehensive experiments on this challenging task demonstrate the effectiveness of each model component, the superiority of our full model over other baselines, and the generalizability of our model to a new VideoQA task. The dataset and codes are open-sourced at: https://github.com/JoseponLee/IntentQA.git
****
LiDAR-UDA: Self-ensembling Through Time for Unsupervised LiDAR Domain Adaptation
Amirreza Shaban, JoonHo Lee, Sanghun Jung, Xiangyun Meng, Byron Boots; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19784-19794
We introduce LiDAR-UDA, a novel two-stage self-training-based Unsupervised Domain Adaptation (UDA) method for LiDAR segmentation. Existing self-training methods use a model trained on labeled source data to generate pseudo labels for target data and refine the predictions via fine-tuning the network on the pseudo labels. These methods suffer from domain shifts caused by different LiDAR sensor configurations in the source and target domains. We propose two techniques to reduce sensor discrepancy and improve pseudo label quality: 1) LiDAR beam subsampling, which simulates different LiDAR scanning patterns by randomly dropping beams; 2) cross-frame ensembling, which exploits temporal consistency of consecutive frames to generate more reliable pseudo labels. Our method is simple, generalizable, and does not incur any extra inference cost. We evaluate our method on several public LiDAR datasets and show that it outperforms the state-of-the-art methods by more than 3.9% mIoU on average for all scenarios. Code will be available at https://github.com/JHLee0513/lidar_uda.
****
Robust Monocular Depth Estimation under Challenging Conditions
Stefano Gasperini, Nils Morbitzer, HyunJun Jung, Nassir Navab, Federico Tombari; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8177-8186
While state-of-the-art monocular depth estimation approaches achieve impressive results in ideal settings, they are highly unreliable under challenging illumination and weather conditions, such as at nighttime or in the presence of rain. In this paper, we uncover these safety-critical issues and tackle them with md4all: a simple and effective solution that works reliably under both adverse and ideal conditions, as well as for different types of learning supervision. We achieve this by exploiting the efficacy of existing methods under perfect settings. Therefore, we provide valid training signals independently of what is in the input. First, we generate a set of complex samples corresponding to the normal training ones. Then, we train the model by guiding its self- or full-supervision by feeding the generated samples and computing the standard losses on the corresponding original images. Doing so enables a single model to recover information across diverse conditions without modifications at inference time. Extensive experime

nts on two challenging public datasets, namely nuScenes and Oxford RobotCar, dem
onstrate the effectiveness of our techniques, outperforming prior works by a lar
ge margin in both standard and challenging conditions. Source code and data are
available at: https://md4all.github.io.
****
Parametric Depth Based Feature Representation Learning for Object Detection and
Segmentation in Bird's-Eye View
Jiayu Yang, Enze Xie, Miaomiao Liu, Jose M. Alvarez; Proceedings of the IEEE/CVF
 International Conference on Computer Vision (ICCV), 2023, pp. 8483-8492
Recent vision-only perception models for autonomous driving achieved promising r
esults by encoding multi-view image features into Bird's-Eye-View (BEV) space. A
 critical step and the main bottleneck of these methods is transforming image fe
atures into the BEV coordinate frame. This paper focuses on leveraging geometry
information, such as depth, to model such feature transformation. Existing works
 rely on non-parametric depth distribution modeling leading to significant memor
y consumption, or ignore the geometry information to address this problem. In co
ntrast, we propose to use parametric depth distribution modeling for feature tra
nsformation. We first lift the 2D image features to the 3D space defined for the
 ego vehicle via a predicted parametric depth distribution for each pixel in eac
h view. Then, we aggregate the 3D feature volume based on the 3D space occupancy
 derived from depth to the BEV frame. Finally, we use the transformed features f
or downstream tasks such as object detection and semantic segmentation. Existing
 semantic segmentation methods do also suffer from an hallucination problem as t
hey do not take visibility information into account. This hallucination can be p
articularly problematic for subsequent modules such as control and planning. To
mitigate the issue, our method provides depth uncertainty and reliable visibilit
y-aware estimations. We further leverage our parametric depth modeling to presen
t a novel visibility-aware evaluation metric that, when taken into account, can
mitigate the hallucination problem. Extensive experiments on object detection an
d semantic segmentation on the nuScenes datasets demonstrate that our method out
performs existing methods on both tasks.
****
MSI: Maximize Support-Set Information for Few-Shot Segmentation
Seonghyeon Moon, Samuel S. Sohn, Honglu Zhou, Sejong Yoon, Vladimir Pavlovic, Mu
hammad Haris Khan, Mubbasir Kapadia; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 19266-19276
FSS (Few-shot segmentation) aims to segment a target class using a small number
of labeled images (support set). To extract the information relevant to target c
lass, a dominant approach in best performing FSS methods removes background feat
ures using a support mask. We observe that this feature excision through a limit
ing support mask introduces an information bottleneck in several challenging FSS
 cases, e.g., for small targets and/or inaccurate target boundaries. To this end
, we present a novel method (MSI), which maximizes the support-set information b
y exploiting two complementary sources of features to generate super correlation
 maps. We validate the effectiveness of our approach by instantiating it into th
ree recent and strong FSS methods. Experimental results on several publicly avai
lable FSS benchmarks show that our proposed method consistently improves perform
ance by visible margins and leads to faster convergence. Our code and trained mo
dels are available at: https://github.com/moonsh/MSI-Maximize-Support-Set-Inform
ation
****
Global Features are All You Need for Image Retrieval and Reranking
Shihao Shao, Kaifeng Chen, Arjun Karpur, Qinghua Cui, André Araujo, Bingyi Cao;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2023, pp. 11036-11046
Image retrieval systems conventionally use a two-stage paradigm, leveraging glob
al features for initial retrieval and local features for reranking. However, the
 scalability of this method is often limited due to the significant storage and
computation cost incurred by local feature matching in the reranking stage. In t
his paper, we present SuperGlobal, a novel approach that exclusively employs glo

bal features for both stages, improving efficiency without sacrificing accuracy. SuperGlobal introduces key enhancements to the retrieval system, specifically focusing on the global feature extraction and reranking processes. For extraction, we identify sub-optimal performance when the widely-used ArcFace loss and Generalized Mean (GeM) pooling methods are combined and propose several new modules to improve GeM pooling. In the reranking stage, we introduce a novel method to update the global features of the query and top-ranked images by only considering feature refinement with a small set of images, thus being very compute and memory efficient. Our experiments demonstrate substantial improvements compared to the state of the art in standard benchmarks. Notably, on the Revisited Oxford+1M Hard dataset, our single-stage results improve by 7.1%, while our two-stage gain reaches 3.7% with a strong 64,865x speedup. Our two-stage system surpasses the current single-stage state-of-the-art by 16.3%, offering a scalable, accurate alternative for high-performing image retrieval systems with minimal time overhead. Code: https://github.com/ShihaoShao-GH/SuperGlobal.

****

DPF-Net: Combining Explicit Shape Priors in Deformable Primitive Field for Unsupervised Structural Reconstruction of 3D Objects

Qingyao Shuai, Chi Zhang, Kaizhi Yang, Xuejin Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14321-14329

Unsupervised methods for reconstructing structures face significant challenges in capturing the geometric details with consistent structures among diverse shapes of the same category. To address this issue, we present a novel unsupervised structural reconstruction method, named DPF-Net, based on a new Deformable Primitive Field (DPF) representation, which allows for high-quality shape reconstruction using parameterized geometric primitives. We design a two-stage shape reconstruction pipeline which consists of a primitive generation module and a primitive deformation module to approximate the target shape of each part progressively. The primitive generation module estimates the explicit orientation, position, and size parameters of parameterized geometric primitives, while the primitive deformation module predicts a dense deformation field based on a parameterized primitive field to recover shape details. The strong shape prior encoded in parameterized geometric primitives enables our DPF-Net to extract high-level structures and recover fine-grained shape details consistently. The experimental results on three categories of objects in diverse shapes demonstrate the effectiveness and generalization ability of our DPF-Net on structural reconstruction and shape segmentation.

****

CORE: Co-planarity Regularized Monocular Geometry Estimation with Weak Supervision

Yuguang Li, Kai Wang, Hui Li, Seon-Min Rhee, Seungju Han, Jihye Kim, Min Yang, Ran Yang, Feng Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8796-8805

The ill-posed nature of monocular 3D geometry (depth map and surface normals) estimation makes it rely mostly on data-driven approaches such as Deep Neural Networks (DNN). However, data acquisition of surface normals, especially the reliable normals, is acknowledged difficult. Commonly, reconstruction of surface normals with high quality is heuristic and time-consuming. Such fact urges methodologies to minimize dependency on ground-truth normals when predicting 3D geometry. In this work, we devise CO-planarity REgularized (CORE) loss functions and Structure-Aware Normal Estimator (SANE). Without involving any knowledge of ground-truth normals, these two designs enable pixel-wise 3D geometry estimation weakly supervised by only ground-truth depth map. For CORE loss functions, the key idea is to exploit locally linear depth-normal orthogonality under spherical coordinates as pixel-level constraints, and utilize our designed Adaptive Polar Regularization (APR) to resolve underlying numerical degeneracies. Meanwhile, SANE easily establishes multi-task learning with CORE loss functions on both depth and surface normal estimation, leading to the whole performance leap. Extensive experiments present the effectiveness of our method on various DNN architectures and dat

a benchmarks. The experimental results demonstrate that our depth estimation achieves the state-of-the-art performance across all metrics on indoor scenes and comparable performance on outdoor scenes. In addition, our surface normal estimation is overall superior.
****
A Sentence Speaks a Thousand Images: Domain Generalization through Distilling CLIP with Language Guidance
Zeyi Huang, Andy Zhou, Zijian Ling, Mu Cai, Haohan Wang, Yong Jae Lee; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11685-11695
Domain generalization studies the problem of training a model with samples from several domains (or distributions) and then testing the model with samples from a new, unseen domain. In this paper, we propose a novel approach for domain generalization that leverages recent advances in large vision-language models, specifically a CLIP teacher model, to train a smaller model that generalizes to unseen domains. The key technical contribution is a new type of regularization that requires the student's learned image representations to be close to the teacher's learned text representations obtained from encoding the corresponding text descriptions of images. We introduce two designs of the loss function, absolute and relative distance, which provide specific guidance on how the training process of the student model should be regularized. We evaluate our proposed method, dubbed RISE (Regularized Invariance with Semantic Embeddings), on various benchmark datasets, and show that it outperforms several state-of-the-art domain generalization methods. To our knowledge, our work is the first to leverage knowledge distillation using a large vision-language model for domain generalization. By incorporating text-based information, RISE improves the generalization capability of machine learning models.
****
H3WB: Human3.6M 3D WholeBody Dataset and Benchmark
Yue Zhu, Nermin Samet, David Picard; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20166-20177
We present a benchmark for 3D human whole-body pose estimation, which involves identifying accurate 3D keypoints on the entire human body, including face, hands, body, and feet. Currently, the lack of a fully annotated and accurate 3D whole-body dataset results in deep networks being trained separately on specific body parts, which are combined during inference. Or they rely on pseudo-groundtruth provided by parametric body models which are not as accurate as detection based methods. To overcome these issues, we introduce the Human3.6M 3D WholeBody (H3WB) dataset, which provides whole-body annotations for the Human3.6M dataset using the COCO Wholebody layout. H3WB comprises 133 whole-body keypoint annotations on 100K images, made possible by our new multi-view pipeline. We also propose three tasks: i) 3D whole-body pose lifting from 2D complete whole-body pose, ii) 3D whole-body pose lifting from 2D incomplete whole-body pose, and iii) 3D whole-body pose estimation from a single RGB image. Additionally, we report several baselines from popular methods for these tasks. Furthermore, we also provide automated 3D whole-body annotations of TotalCapture and experimentally show that when used with H3WB it helps to improve the performance.
****
Yes, we CANN: Constrained Approximate Nearest Neighbors for Local Feature-Based Visual Localization
Dror Aiger, Andre Araujo, Simon Lynen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13339-13349
Large-scale visual localization systems continue to relyon 3D point clouds built from image collections usingstructure-from-motion. While the 3D points in these modelsare represented using local image features, directly match-ing a query image's local features against the point cloud ischallenging due to the scale of the nearest-neighbor searchproblem. Many recent approaches to visual localization havethus proposed a hybrid method, where first a global (per im-age) embedding is used to retrieve a small subset of databaseimages, and local features of the query are matched onlyagainst those. It seems to have become common belief thatg

lobal embeddings are critical for said image-retrieval invisual localization, despite the significant downside of hav-ing to compute two feature types for each query image. Inthis paper, we take a step back from this assumption and pro-pose Constrained Approximate Nearest Neighbors (CANN),a joint solution of k-nearest-neighbors across both the ge-ometry and appearance space using only local features. Wefirst derive the theoretical foundation for k-nearest-neighborretrieval across multiple metrics and then showcase howCANN improves visual localization. Our experiments onpublic localization benchmarks demonstrate that our methodsignificantly outperforms both state-of-the-art global feature-based retrieval and approaches using local feature aggrega-tion schemes. Moreover, it is an order of magnitude faster inboth index and query time than feature aggregation schemesfor these datasets. Code will be released.
****
Multi-Object Navigation with Dynamically Learned Neural Implicit Representations
Pierre Marza, Laetitia Matignon, Olivier Simonin, Christian Wolf; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11004-11015
Understanding and mapping a new environment are core abilities of any autonomously navigating agent. While classical robotics usually estimates maps in a stand-alone manner with SLAM variants, which maintain a topological or metric representation, end-to-end learning of navigation keeps some form of memory in a neural network. Networks are typically imbued with inductive biases, which can range from vectorial representations to birds-eye metric tensors or topological structures. In this work, we propose to structure neural networks with two neural implicit representations, which are learned dynamically during each episode and map the content of the scene: (i) the Semantic Finder predicts the position of a previously seen queried object; (ii) the Occupancy and Exploration Implicit Representation encapsulates information about explored area and obstacles, and is queried with a novel global read mechanism which directly maps from function space to a usable embedding space. Both representations are leveraged by an agent trained with Reinforcement Learning (RL) and learned online during each episode. We evaluate the agent on Multi-Object Navigation and show the high impact of using neural implicit representations as a memory source.
****
NPC: Neural Point Characters from Video
Shih-Yang Su, Timur Bagautdinov, Helge Rhodin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14795-14805
High-fidelity human 3D models can now be learned directly from videos, typically by combining a template-based surface model with neural representations. However, obtaining a template surface requires expensive multi-view capture systems, laser scans, or strictly controlled conditions. Previous methods avoid using a template but rely on a costly or ill-posed mapping from observation to canonical space. We propose a hybrid point-based representation for animatable humans that does not require an explicit surface model, while being generalizable to novel poses. For a given video, our method automatically produces an explicit set of 3D points representing approximate canonical geometry, and learns an articulated deformation model that produces pose-dependent point transformations. The points serve both as a scaffold for high-frequency neural features and an anchor for efficiently mapping between observation and canonical space. We demonstrate on established benchmarks that our representation overcomes limitations of prior work operating in either canonical or in observation space. Moreover, our automatic point extraction approach enables learning models of human and animal characters alike, matching the performance of the methods using rigged surface templates despite being more general. Project website: https: //lemonatsu.github.io/npc/.
****
LDP-Feat: Image Features with Local Differential Privacy
Francesco Pittaluga, Bingbing Zhuang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17580-17590
Modern computer vision services often require users to share raw feature descriptors with an untrusted server. This presents an inherent privacy risk, as raw de

scriptors may be used to recover the source images from which they were extracte d. To address this issue, researchers recently proposed privatizing image featur es by embedding them within an affine subspace containing the original feature a s well as adversarial feature samples. In this paper, we propose two novel inver sion attacks to show that it is possible to (approximately) recover the original image features from these embeddings, allowing us to recover privacy-critical i mage content. In light of such successes and the lack of theoretical privacy gua rantees afforded by existing visual privacy methods, we further propose the firs t method to privatize image features via local differential privacy, which, unli ke prior approaches, provides a guaranteed bound for privacy leakage regardless of the strength of the attacks. In addition, our method yields strong performanc e in visual localization as a downstream task while enjoying the privacy guarant ee.
****
Pre-Training-Free Image Manipulation Localization through Non-Mutually Exclusive Contrastive Learning
Jizhe Zhou, Xiaochen Ma, Xia Du, Ahmed Y. Alhammadi, Wentao Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22346-22356
Deep Image Manipulation Localization (IML) models suffer from training data insu fficiency and thus heavily rely on pre-training. We argue that contrastive learn ing is more suitable to tackle the data insufficiency problem for IML. Crafting mutually exclusive positives and negatives is the prerequisite for contrastive l earning. However, when adopting contrastive learning in IML, we encounter three categories of image patches: tampered, authentic, and contour patches. Tampered and authentic patches are naturally mutually exclusive, but contour patches cont aining both tampered and authentic pixels are non-mutually exclusive to them. Si mply abnegating these contour patches results in a drastic performance loss sinc e contour patches are decisive to the learning outcomes. Hence, we propose the N on-mutually exclusive Contrastive Learning (NCL) framework to rescue conventiona l contrastive learning from the above dilemma. In NCL, to cope with the non-mutu ally exclusivity, we first establish a pivot structure with dual branches to con stantly switch the role of contour patches between positives and negatives while training. Then, we devise a pivot-consistent loss to avoid spatial corruption c aused by the role-switching process. In this manner, NCL both inherits the self-supervised merits to address the data insufficiency and retains a high manipulat ion localization accuracy. Extensive experiments verify that our NCL achieves st ate-of-the-art performance on all five benchmarks without any pre-training and i s more robust on unseen real-life samples. https://github.com/Knightzjz/NCL-IML
****
MRN: Multiplexed Routing Network for Incremental Multilingual Text Recognition
Tianlun Zheng, Zhineng Chen, Bingchen Huang, Wei Zhang, Yu-Gang Jiang; Proceedin gs of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18644-18653
Multilingual text recognition (MLTR) systems typically focus on a fixed set of l anguages, which makes it difficult to handle newly added languages or adapt to e ver-changing data distribution. In this paper, we propose the Incremental MLTR (IMLTR) task in the context of incremental learning (IL), where different languag es are introduced in batches. IMLTR is particularly challenging due to rehearsal -imbalance, which refers to the uneven distribution of sample characters in the rehearsal set, used to retain a small amount of old data as past memories. To ad dress this issue, we propose a Multiplexed Routing Network (MRN). MRN trains a r ecognizer for each language that is currently seen. Subsequently, a language dom ain predictor is learned based on the rehearsal set to weigh the recognizers. Si nce the recognizers are derived from the original data, MRN effectively reduces the reliance on older data and better fights against catastrophic forgetting, th e core issue in IL. We extensively evaluate MRN on MLT17 and MLT19 datasets. It outperforms existing general-purpose IL methods by large margins, with average a ccuracy improvements ranging from 10.3% to 35.8% under different settings. Code is available at https://github.com/simplify23/MRN.

****
Domain Generalization Guided by Gradient Signal to Noise Ratio of Parameters
Mateusz Michalkiewicz, Masoud Faraki, Xiang Yu, Manmohan Chandraker, Mahsa Baktashmotlagh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6177-6188
Overfitting to the source domain is a common issue in gradient-based training of deep neural networks. To compensate for the over-parameterized models, numerous regularization techniques have been introduced such as those based on dropout. While these methods achieve significant improvements on classical benchmarks such as ImageNet, their performance diminishes with the introduction of domain shift in the test set i.e. when the unseen data comes from a significantly different distribution. In this paper, we move away from the classical approach of Bernoulli sampled dropout mask construction and propose to base the selection on gradient-signal-to-noise ratio (GSNR) of network's parameters. Specifically, at each training step, parameters with high GSNR will be discarded. Furthermore, we alleviate the burden of manually searching for the optimal dropout ratio by leveraging a meta-learning approach. We evaluate our method on standard domain generalization benchmarks and achieve competitive results on classification and face anti-spoofing problems.
****
Counterfactual-based Saliency Map: Towards Visual Contrastive Explanations for Neural Networks
Xue Wang, Zhibo Wang, Haiqin Weng, Hengchang Guo, Zhifei Zhang, Lu Jin, Tao Wei, Kui Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2042-2051
Explaining deep models in a human-understandable way has been explored by many works that mostly explain why an input causes a corresponding prediction (ie., Why P?). However, seldom they could handle those more complex causal questions like "why P rather than Q?" and "why one is P while another is Q?", which would better help humans understand the behavior of deep models. Considering the insufficient study on such complex causal questions, we make the first attempt to explain different causal questions by contrastive explanations in a unified framework, ie., Counterfactual Contrastive Explanation (CCE), which visually and intuitively explains the aforementioned questions via a novel positive-negative saliency-based explanation scheme. More specifically, we propose a content-aware counterfactual perturbing algorithm to stimulate contrastive examples, from which a pair of positive and negative saliency maps could be derived to contrastively explain why P (positive class) rather than Q (negative class). Beyond existing works, our counterfactual perturbation meets the principles of validity, sparsity, and data distribution closeness at the same time. In addition, by slightly adjusting the objective of perturbation, our framework can adapt to different causal questions. Extensive experimental evaluation demonstrates the effectiveness and superior performance of the proposed CCE on different benchmark metrics for interpretability, including Sanity Check, Class Deviation Score and Insertion-Deletion tests. A user study is conducted and the results show that user confidence is increasing significantly when presented with CCE compared to standard saliency map baselines.
****
MST-compression: Compressing and Accelerating Binary Neural Networks with Minimum Spanning Tree
Quang Hieu Vo, Linh-Tam Tran, Sung-Ho Bae, Lok-Won Kim, Choong Seon Hong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6091-6100
Binary neural networks (BNNs) have been widely adopted to reduce the computational cost and memory storage on edge-computing devices by using one bit representation for activations and weights. However, as neural networks become wider/deeper to improve accuracy and meet practical requirements, the computational burden remains a significant challenge even on the binary version. To address these issues, this paper proposes a novel method called Minimum Spanning Tree (MST) compression that learns to compress and accelerate BNNs. The proposed architecture le

verages an observation from previous works that an output channel in a binary convolution can be computed using another output channel and XNOR operations with weights that differ from the weights of the reused channel. We first construct a fully connected graph with vertices corresponding to output channels, where the distance between two vertices is the number of different values between the weight sets used for these outputs. Then, the MST of the graph with the minimum depth is proposed to reorder output calculations, aiming to reduce computational cost and latency. Moreover, we propose a new learning algorithm to reduce the total MST distance during training. Experimental results on benchmark models demonstrate that our method achieves significant compression ratios with negligible accuracy drops, making it a promising approach for resource-constrained edge-computing devices.
****
MOST: Multiple Object Localization with Self-Supervised Transformers for Object Discovery
Sai Saketh Rambhatla, Ishan Misra, Rama Chellappa, Abhinav Shrivastava; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15823-15834
We tackle the challenging task of unsupervised object localization in this work. Recently, transformers trained with self-supervised learning have been shown to exhibit object localization properties without being trained for this task. In this work, we present Multiple Object localization with Self-supervised Transformers (MOST) that uses features of transformers trained using self-supervised learning to localize multiple objects in real world images. MOST analyzes the similarity maps of the features using box counting; a fractal analysis tool to identify tokens lying on foreground patches. The identified tokens are then clustered together, and tokens of each cluster are used to generate bounding boxes on foreground regions. Unlike recent state-of-the-art object localization methods, MOST can localize multiple objects per image and outperforms SOTA algorithms on several object localization and discovery benchmarks on PASCAL-VOC 07, 12 and COCO20k datasets. Additionally, we show that MOST can be used for self-supervised pretraining of object detectors, and yields consistent improvements on fully, semi-supervised object detection and unsupervised region proposal generation.Our project is publicly available at rssaketh.github.io/most.
****
IIEU: Rethinking Neural Feature Activation from Decision-Making
Sudong Cai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5796-5806
Nonlinear Activation (Act) models which help fit the underlying mappings are critical for neural representation learning. Neuronal behaviors inspire basic Act functions, e.g., Softplus and ReLU. We instead seek improved explainable Act models by re-interpreting neural feature Act from a new philosophical perspective of Multi-Criteria Decision-Making (MCDM). By treating activation models as selective feature re-calibrators that suppress/emphasize features according to their importance scores measured by feature-filter similarities, we propose a set of specific properties of effective Act models with new intuitions. This helps us identify the unexcavated yet critical problem of mismatched feature scoring led by the differentiated norms of the features and filters. We present the Instantaneous Importance Estimation Units (IIEUs), a novel class of interpretable Act models that address the problem by re-calibrating the feature with the Instantaneous Importance (II) score (which we refer to as) estimated with the adaptive norm-decoupled feature-filter similarities, capable of modeling the cross-layer and -channel cues at a low cost. The extensive experiments on various vision benchmarks demonstrate the significant improvements of our IIEUs over the SOTA Act models and validate our interpretation of feature Act. By replacing the popular/SOTA Act models with IIEUs, the small ResNet-26s outperform/match the large ResNet-101s on ImageNet with far fewer parameters and computations.
****
Integrally Migrating Pre-trained Transformer Encoder-decoders for Visual Object Detection

Feng Liu, Xiaosong Zhang, Zhiliang Peng, Zonghao Guo, Fang Wan, Xiangyang Ji, Qixiang Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6825-6834

Modern object detectors have taken the advantages of backbone networks pre-trained on large scale datasets. Except for the backbone networks, however, other components such as the detector head and the feature pyramid network (FPN) remain trained from scratch, which hinders the generalization capacity of detectors. In this study, we propose to integrally migrate pre-trained transformer encoder-decoders (imTED) to a detector, constructing a feature extraction path which is "fully pre-trained" so that detectors' generalization capacity is maximized. The essential differences between imTED with the baseline detector are twofold: (1) migrating the pre-trained transformer decoder to the detector head while removing the randomly initialized FPN from the feature extraction path; and (2) defining a multi-scale feature modulator (MFM) to enhance scale adaptability. Such designs not only reduce randomly initialized parameters significantly but also unify detector training with representation learning intendedly. Experiments on the MS COCO object detection dataset show that imTED consistently outperforms its counterparts by 2.4 AP. Without bells and whistles, imTED improves the state-of-the-art of few-shot object detection by up to 7.6 AP. Code is released at https://github.com/LiewFeng/imTED.

****

V-FUSE: Volumetric Depth Map Fusion with Long-Range Constraints

Nathaniel Burgdorfer, Philippos Mordohai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3449-3458

We introduce a learning-based depth map fusion framework that accepts a set of depth and confidence maps generated by a Multi-View Stereo (MVS) algorithm as input and improves them. This is accomplished by integrating volumetric visibility constraints that encode long-range surface relationships across different views into an end-to-end trainable architecture. We also introduce a depth search window estimation sub-network trained jointly with the larger fusion sub-network to reduce the depth hypothesis search space along each ray. Our method learns to model depth consensus and violations of visibility constraints directly from the data; effectively removing the necessity of fine-tuning fusion parameters. Extensive experiments on MVS datasets show substantial improvements in the accuracy of the output fused depth and confidence maps.

****

CrossLoc3D: Aerial-Ground Cross-Source 3D Place Recognition

Tianrui Guan, Aswath Muthuselvam, Montana Hoover, Xijun Wang, Jing Liang, Adarsh Jagan Sathyamoorthy, Damon Conover, Dinesh Manocha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11335-11344

We present CrossLoc3D, a novel 3D place recognition method that solves a large-scale point matching problem in a cross-source setting. Cross-source point cloud data corresponds to point sets captured by depth sensors with different accuracies or from different distances and perspectives. We address the challenges in terms of developing 3D place recognition methods that account for the representation gap between points captured by different sources. Our method handles cross-source data by utilizing multi-grained features and selecting convolution kernel sizes that correspond to most prominent features. Inspired by the diffusion models, our method uses a novel iterative refinement process that gradually shifts the embedding spaces from different sources to a single canonical space for better metric learning. In addition, we present CS-Campus3D, the first 3D aerial-ground cross-source dataset consisting of point cloud data from both aerial and ground LiDAR scans. The point clouds in CS-Campus3D have representation gaps and other features like different views, point densities, and noise patterns. We show that our CrossLoc3D algorithm can achieve an improvement of 4.74% - 15.37% in terms of the top 1 average recall on our CS-Campus3D benchmark and achieves performance comparable to state-of-the-art 3D place recognition method on the Oxford RobotCar. We will release the code and CS-Campus3D benchmark.

****

Recursive Video Lane Detection

Dongkwon Jin, Dahyun Kim, Chang-Su Kim; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2023, pp. 8473-8482
A novel algorithm to detect road lanes in videos, called recursive video lane de
tector (RVLD), is proposed in this paper, which propagates the state of a curren
t frame recursively to the next frame. RVLD consists of an intra-frame lane dete
ctor (ILD) and a predictive lane detector (PLD). First, we design ILD to localiz
e lanes in a still frame. Second, we develop PLD to exploit the information of t
he previous frame for lane detection in a current frame. To this end, we estimat
e a motion field and warp the previous output to the current frame. Using the wa
rped information, we refine the feature map of the current frame to detect lanes
 more reliably. Experimental results show that RVLD outperforms existing detecto
rs on video lane datasets. Our codes are available at https://github.com/dongkwo
njin/RVLD.
****
GECCO: Geometrically-Conditioned Point Diffusion Models
Micha■ J Tyszkiewicz, Pascal Fua, Eduard Trulls; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 2128-2138
Diffusion models generating images conditionally on text, such as Dall-E 2 and S
table Diffusion, have recently made a splash far beyond the computer vision comm
unity. Here, we tackle the related problem of generating point clouds, both unco
nditionally, and conditionally with images. For the latter, we introduce a novel
 geometrically-motivated conditioning scheme based on projecting sparse image fe
atures into the point cloud and attaching them to each individual point, at ever
y step in the denoising process. This approach improves geometric consistency an
d yields greater fidelity than current methods relying on unstructured, global l
atent codes. Additionally, we show how to apply recent continuous-time diffusion
 schemes. Our method performs on par or above the state of art on conditional an
d unconditional experiments on synthetic data, while being faster, lighter, and
delivering tractable likelihoods. We show it can also scale to diverse indoors s
cenes.
****
Unsupervised Self-Driving Attention Prediction via Uncertainty Mining and Knowle
dge Embedding
Pengfei Zhu, Mengshi Qi, Xia Li, Weijian Li, Huadong Ma; Proceedings of the IEEE
/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8558-8568
Predicting attention regions of interest is an important yet challenging task fo
r self-driving systems. Existing methodologies rely on large-scale labeled traff
ic datasets that are labor-intensive to obtain. Besides, the huge domain gap bet
ween natural scenes and traffic scenes in current datasets also limits the poten
tial for model training. To address these challenges, we are the first to introd
uce an unsupervised way to predict self-driving attention by uncertainty modelin
g and driving knowledge integration. Our approach's Uncertainty Mining Branch (U
MB) discovers commonalities and differences from multiple generated pseudo-label
s achieved from models pre-trained on natural scenes by actively measuring the u
ncertainty. Meanwhile, our Knowledge Embedding Block (KEB) bridges the domain ga
p by incorporating driving knowledge to adaptively refine the generated pseudo-l
abels. Quantitative and qualitative results with equivalent or even more impress
ive performance compared to fully-supervised state-of-the-art approaches across
all three public datasets demonstrate the effectiveness of the proposed method a
nd the potential of this direction. The code is available at https://github.com/
zaplm/DriverAttention.
****
PETRv2: A Unified Framework for 3D Perception from Multi-Camera Images
Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, Xiangyu Zh
ang; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 3262-3272
In this paper, we propose PETRv2, a unified framework for 3D perception from mul
ti-view images. Based on PETR, PETRv2 explores the effectiveness of temporal mod
eling, which utilizes the temporal information of previous frames to boost 3D ob
ject detection. More specifically, we extend the 3D position embedding (3D PE) i

n PETR for temporal modeling. The 3D PE achieves the temporal alignment on object position of different frames. To support for multi-task learning (e.g., BEV segmentation and 3D lane detection), PETRv2 provides a simple yet effective solution by introducing task-specific queries, which are initialized under different spaces. PETRv2 achieves state-of-the-art performance on 3D object detection, BEV segmentation and 3D lane detection. Detailed robustness analysis is also conducted on PETR framework. Code is available at https://github.com/megvii-research/PETR.
****
Out-of-Domain GAN Inversion via Invertibility Decomposition for Photo-Realistic Human Face Manipulation
Xin Yang, Xiaogang XU, Yingcong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7492-7501
The fidelity of Generative Adversarial Networks (GAN) inversion is impeded by Out-Of-Domain (OOD) areas (e.g., background, accessories) in the image.
 Detecting the OOD areas beyond the generation ability of the pre-trained model and blending these regions with the input image can enhance fidelity. The "invertibility mask" figures out these OOD areas, and existing methods predict the mask with the reconstruction error. However, the estimated mask is usually inaccurate due to the influence of the reconstruction error in the In-Domain (ID) area. In this paper, we propose a novel framework that enhances the fidelity of human face inversion by designing a new module to decompose the input images to ID and OOD partitions with invertibility masks. Unlike previous works, our invertibility detector is simultaneously learned with a spatial alignment module. We iteratively align the generated features to the input geometry and reduce the reconstruction error in the ID regions. Thus, the OOD areas are more distinguishable and can be precisely predicted. Then, we improve the fidelity of our results by blending the OOD areas from the input image with the ID GAN inversion results.
 Our method produces photo-realistic results for real-world human face image inversion and manipulation. Extensive experiments demonstrate our method's superiority over existing methods in the quality of GAN inversion and attribute manipulation.
****
SAFE: Machine Unlearning With Shard Graphs
Yonatan Dukler, Benjamin Bowman, Alessandro Achille, Aditya Golatkar, Ashwin Swaminathan, Stefano Soatto; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17108-17118
We present Synergy Aware Forgetting Ensemble (SAFE), a method to adapt large models on a diverse collection of data while minimizing the expected cost to remove the influence of training samples from the trained model. This process, also known as selective forgetting or unlearning, is often conducted by partitioning a dataset into shards, training fully independent models on each, then ensembling the resulting models. Increasing the number of shards reduces the expected cost to forget but at the same time it increases inference cost and reduces the final accuracy of the model since synergistic information between samples is lost during the independent model training. Rather than treating each shard as independent, SAFE introduces the notion of a shard graph, which allows incorporating limited information from other shards during training, trading off a modest increase in expected forgetting cost with a significant increase in accuracy, all while still attaining complete removal of residual influence after forgetting. SAFE uses a lightweight system of adapters which can be trained while reusing most of the computations. This allows SAFE to be trained on shards an order-of-magnitude smaller than current state-of-the-art methods (thus reducing the forgetting costs) while also maintaining high accuracy, as we demonstrate empirically on fine-grained computer vision datasets.
****
Learning Trajectory-Word Alignments for Video-Language Tasks
Xu Yang, Zhangzikang Li, Haiyang Xu, Hanwang Zhang, Qinghao Ye, Chenliang Li, Ming Yan, Yu Zhang, Fei Huang, Songfang Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2504-2514

In a video, an object usually appears as the trajectory, i.e., it spans over a few spatial but longer temporal patches, that contains abundant spatiotemporal contexts. However, modern Video-Language BERTs (VDL-BERTs) neglect this trajectory characteristic that they usually follow image-language BERTs (IL-BERTs) to deploy the patch-to-word (P2W) attention that may over-exploit trivial spatial contexts and neglect significant temporal contexts. To amend this, we propose a novel TW-BERT to learn Trajectory-Word alignment by a newly designed trajectory-to-word (T2W) attention for solving video-language tasks. Moreover, previous VDL-BERTs usually uniformly sample a few frames into the model while different trajectories have diverse graininess, i.e., some trajectories span longer frames and some span shorter, and using a few frames will lose certain useful temporal contexts. However, simply sampling more frames will also make pre-training infeasible due to the largely increased training burdens. To alleviate the problem, during the fine-tuning stage, we insert a novel Hierarchical Frame-Selector (HFS) module into the video encoder. HFS gradually selects the suitable frames conditioned on the text context for the later cross-modal encoder to learn better trajectory-word alignments. By the proposed T2W attention and HFS, our TW-BERT achieves SOTA performances on text-to-video retrieval tasks, and comparable performances on video question-answering tasks with some VDL-BERTs trained on much more data. The code will be available in the supplementary material.
****
OrthoPlanes: A Novel Representation for Better 3D-Awareness of GANs
Honglin He, Zhuoqian Yang, Shikai Li, Bo Dai, Wayne Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22996-23007
We present a new method for generating realistic and view-consistent images with fine geometry from 2D image collections. Our method proposes a hybrid explicit-implicit representation called OrthoPlanes, which encodes fine-grained 3D information in feature maps that can be efficiently generated by modifying 2D StyleGANs. Compared to previous representations, our method has better scalability and expressiveness with clear and explicit information. As a result, our method can handle more challenging view-angles and synthesize articulated objects with high spatial degree of freedom. Experiments demonstrate that our method achieves state-of-the-art results on FFHQ and SHHQ datasets, both quantitatively and qualitatively.
****
Geometry-guided Feature Learning and Fusion for Indoor Scene Reconstruction
Ruihong Yin, Sezer Karaoglu, Theo Gevers; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3652-3661
In addition to color and textual information, geometry provides important cues for 3D scene reconstruction. However, current reconstruction methods only include geometry at the feature level thus not fully exploiting the geometric information. In contrast, this paper proposes a novel geometry integration mechanism for 3D scene reconstruction. Our approach incorporates 3D geometry at three levels, i.e. feature learning, feature fusion, and network supervision. First, geometry-guided feature learning encodes geometric priors to contain view-dependent information. Second, a geometry-guided adaptive feature fusion is introduced which utilizes the geometric priors as a guidance to adaptively generate weights for multiple views. Third, at the supervision level, taking the consistency between 2D and 3D normals into account, a consistent 3D normal loss is designed to add local constraints. Large-scale experiments are conducted on the ScanNet dataset, showing that volumetric methods with our geometry integration mechanism outperform state-of-the-art methods quantitatively as well as qualitatively. Volumetric methods with ours also show good generalization on the 7-Scenes and TUM RGB-D datasets.
****
Atmospheric Transmission and Thermal Inertia Induced Blind Road Segmentation with a Large-Scale Dataset TBRSD
Junzhang Chen, Xiangzhi Bai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1053-1063
Computer vision-based walking assistants are prominent tools for aiding visually

impaired people in navigation. Blind road segmentation is a key element in these walking assistant systems. However, most walking assistant systems rely on visual light images, which is dangerous in weak illumination environments such as darkness or fog. To address this issue and enhance the safety of vision-based walking assistant systems, we developed a thermal infrared blind road segmentation neural network (TINN). In contrast to conventional segmentation techniques that primarily concentrate on enhancing feature extraction and perception, our approach is geared towards preserving the inherent radiation characteristics within the thermal imaging process. Initially, we modelled two critical factors in thermal infrared imaging - thermal light atmospheric transmission and thermal inertia effect. Subsequently, we use an encoder-decoder architecture to fuse the feathers extracted by the two modules. Additionally, to train the network and evaluate the effectiveness of the proposed method, we constructed a large-scale thermal infrared blind road segmentation dataset named TBRSD consists 5180 pixel-level manual annotations. The experimental results demonstrate that our method outperforms existing techniques and achieves state-of-the-art performance in thermal blind road segmentation, as validated on benchmark thermal infrared semantic segmentation datasets such as MFNet and SODA. The dataset and our code are both publicly available in https://github.com/chenjzBUAA/TBRSD or http://xzbai.buaa.edu.cn/datasets.html.
****
NeTO:Neural Reconstruction of Transparent Objects with Self-Occlusion Aware Refraction-Tracing
Zongcheng Li, Xiaoxiao Long, Yusen Wang, Tuo Cao, Wenping Wang, Fei Luo, Chunxia Xiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18547-18557
We present a novel method called NeTO, for capturing the 3D geometry of solid transparent objects from 2D images via volume rendering. Reconstructing transparent objects is a very challenging task, which is ill-suited for general-purpose reconstruction techniques due to the specular light transport phenomena. Although existing refraction-tracing-based methods, designed especially for this task, achieve impressive results, they still suffer from unstable optimization and loss of fine details since the explicit surface representation they adopted is difficult to be optimized, and the self-occlusion problem is ignored for refraction-tracing. In this paper, we propose to leverage implicit Signed Distance Function (SDF) as surface representation and optimize the SDF field via volume rendering with a self-occlusion aware refractive ray tracing. The implicit representation enables our method to be capable of reconstructing high-quality reconstruction even with a limited set of views, and the self-occlusion aware strategy makes it possible for our method to accurately reconstruct the self-occluded regions. Experiments show that our method achieves faithful reconstruction results and outperforms prior works by a large margin. Visit our project page at https://www.xxlong.site/NeTO/.
****
Boosting 3-DoF Ground-to-Satellite Camera Localization Accuracy via Geometry-Guided Cross-View Transformer
Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, Hongdong Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21516-21526
Image retrieval-based cross-view localization methods often lead to very coarse camera pose estimation, due to the limited sampling density of the database satellite images. In this paper, we propose a method to increase the accuracy of a ground camera's location and orientation by estimating the relative rotation and translation between the ground-level image and its matched/retrieved satellite image.
Our approach designs a geometry-guided cross-view transformer that combines the benefits of conventional geometry and learnable cross-view transformers to map the ground-view observations to an overhead view.
Given the synthesized overhead view and observed satellite feature maps, we construct a neural pose optimizer with strong global information embedding ability

to estimate the relative rotation between them. After aligning their rotations, we develop an uncertainty-guided spatial correlation to generate a probability map of the vehicle locations, from which the relative translation can be determined.
Experimental results demonstrate that our method significantly outperforms the state-of-the-art. Notably, the likelihood of restricting the vehicle lateral pose to be within 1m of its Ground Truth (GT) value on the cross-view KITTI dataset has been improved from 35.54% to 76.44%, and the likelihood of restricting the vehicle orientation to be within 1 degree of its GT value has been improved from 19.64% to 99.10%.
****
Efficient-VQGAN: Towards High-Resolution Image Generation with Efficient Vision Transformers
Shiyue Cao, Yueqin Yin, Lianghua Huang, Yu Liu, Xin Zhao, Deli Zhao, Kaigi Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7368-7377
Vector-quantized image modeling has shown great potential in synthesizing high-quality images. However, generating high-resolution images remains a challenging task due to the quadratic computational overhead of the self-attention process. In this study, we seek to explore a more efficient two-stage framework for high-resolution image generation with improvements in the following three aspects. (1) Based on the observation that the first quantization stage has solid local property, we employ a local attention-based quantization model instead of the global attention mechanism used in previous methods, leading to better efficiency and reconstruction quality. (2) We emphasize the importance of multi-grained feature interaction during image generation and introduce an efficient attention mechanism that combines global attention (long-range semantic consistency within the whole image) and local attention (fined-grained details). This approach results in faster generation speed, higher generation fidelity, and improved resolution. (3) We propose a new generation pipeline incorporating autoencoding training and autoregressive generation strategy, demonstrating a better paradigm for image synthesis. Extensive experiments demonstrate the superiority of our approach in high-quality and high-resolution image reconstruction and generation.
****
DLGSANet: Lightweight Dynamic Local and Global Self-Attention Networks for Image Super-Resolution
Xiang Li, Jiangxin Dong, Jinhui Tang, Jinshan Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12792-12801
We propose an effective lightweight dynamic local and global self-attention network (DLGSANet) to solve image super-resolution. Our method explores the properties of Transformers while having low computational costs. Motivated by the network designs of Transformers, we develop a simple yet effective multi-head dynamic local self-attention (MHDLSA) module to extract local features efficiently. In addition, we note that existing Transformers usually explore all similarities of the tokens between the queries and keys for the feature aggregation. However, not all the tokens from the queries are relevant to those in keys, using all the similarities does not effectively facilitate the high-resolution image reconstruction. To overcome this problem, we develop a sparse global self-attention (SparseGSA) module to select the most useful similarity values so that the most useful global features can be better utilized for the high-resolution image reconstruction. We develop a hybrid dynamic-Transformer block (HDTB) that integrates the MHDLSA and SparseGSA for both local and global feature exploration. To ease the network training, we formulate the HDTBs into a residual hybrid dynamic-Transformer group (RHDTG). By embedding the RHDTGs into an end-to-end trainable network, we show that our proposed method has fewer network parameters and lower computational costs while achieving competitive performance against state-of-the-art ones in terms of accuracy. More information is available at https://neonleexiang.github.io/DLGSANet/.
****
Adaptive Reordering Sampler with Neurally Guided MAGSAC

Tong Wei, Jiri Matas, Daniel Barath; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18163-18173
We propose a new sampler for robust estimators that always selects the sample with the highest probability of consisting only of inliers. After every unsuccessful iteration, the inlier probabilities are updated in a principled way via a Bayesian approach. The probabilities obtained by the deep network are used as prior (so-called neural guidance) inside the sampler. Moreover, we introduce a new loss that exploits, in a geometrically justifiable manner, the orientation and scale that can be estimated for any type of feature, e.g., SIFT or SuperPoint, to estimate two-view geometry. The new loss helps to learn higher-order information about the underlying scene geometry. Benefiting from the new sampler and the proposed loss, we combine the neural guidance with the state-of-the-art MAGSAC++. Adaptive Reordering Sampler with Neurally Guided MAGSAC (ARS-MAGSAC) is superior to the state-of-the-art in terms of accuracy and run-time on the PhotoTourism and KITTI datasets for essential and fundamental matrix estimation. The code and trained models are available at https://github.com/weitong8591/ars_magsac.
****
Learning Cross-Representation Affinity Consistency for Sparsely Supervised Biomedical Instance Segmentation
Xiaoyu Liu, Wei Huang, Zhiwei Xiong, Shenglong Zhou, Yueyi Zhang, Xuejin Chen, Zheng-Jun Zha, Feng Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21107-21117
Sparse instance-level supervision has recently been explored to address insufficient annotation in biomedical instance segmentation, which is easier to annotate crowded instances and better preserves instance completeness for 3D volumetric datasets compared to common semi-supervision.In this paper, we propose a sparsely supervised biomedical instance segmentation framework via cross-representation affinity consistency regularization. Specifically, we adopt two individual networks to enforce the perturbation consistency between an explicit affinity map and an implicit affinity map to capture both feature-level instance discrimination and pixel-level instance boundary structure. We then select the highly confident region of each affinity map as the pseudo label to supervise the other one for affinity consistency learning. To obtain the highly confident region, we propose a pseudo-label noise filtering scheme by integrating two entropy-based decision strategies. Extensive experiments on four biomedical datasets with sparse instance annotations show the state-of-the-art performance of our proposed framework. For the first time, we demonstrate the superiority of sparse instance-level supervision on 3D volumetric datasets, compared to common semi-supervision under the same annotation cost.
****
Black-Box Unsupervised Domain Adaptation with Bi-Directional Atkinson-Shiffrin Memory
Jingyi Zhang, Jiaxing Huang, Xueying Jiang, Shijian Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11771-11782
Black-box unsupervised domain adaptation (UDA) learns with source predictions of target data without accessing either source data or source models during training, and it has clear superiority in data privacy and flexibility in target network selection. However, the source predictions of target data are often noisy and training with them is prone to learning collapses. We propose BiMem, a bi-directional memorization mechanism that learns to remember useful and representative information to correct noisy pseudo labels on the fly, leading to robust black-box UDA that can generalize across different visual recognition tasks. BiMem constructs three types of memory, including sensory memory, short-term memory, and long-term memory, which interact in a bi-directional manner for comprehensive and robust memorization of learnt features. It includes a forward memorization flow that identifies and stores useful features and a backward calibration flow that rectifies features' pseudo labels progressively. Extensive experiments show that BiMem achieves superior domain adaptation performance consistently across various visual recognition tasks such as image classification, semantic segmentation and object detection.

****
Towards Fair and Comprehensive Comparisons for Image-Based 3D Object Detection
Xinzhu Ma, Yongtao Wang, Yinmin Zhang, Zhiyi Xia, Yuan Meng, Zhihui Wang, Haojie Li, Wanli Ouyang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6425-6435
In this work, we build a modular-designed codebase, formulate strong training recipes, design an error diagnosis toolbox, and discuss current methods for image-based 3D object detection. Specifically, different from other highly mature tasks, e.g., 2D object detection, the community of image-based 3D object detection is still evolving, where methods often adopt different training recipes and tricks resulting in unfair evaluations and comparisons. What is worse, these tricks may overwhelm their proposed designs in performance, even leading to wrong conclusions. To address this issue, we build a module-designed codebase and formulate unified training standards for the community. Furthermore, we also design an error diagnosis toolbox to measure the detailed characterization of detection models. Using these tools, we analyze current methods in-depth under varying settings and provide discussions for some open questions, e.g., discrepancies in conclusions on KITTI-3D and nuScenes datasets, which have led to different dominant methods for these datasets. We hope that this work will facilitate future research in vision-based 3D detection. Our codes will be released at https://github.com/OpenGVLab/3dodi.
****
Disentangling Spatial and Temporal Learning for Efficient Image-to-Video Transfer Learning
Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, Nong Sang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13934-13944
Recently, large-scale pre-trained language-image models like CLIP have shown extraordinary capabilities for understanding spatial contents, but naively transferring such models to video recognition still suffers from unsatisfactory temporal modelling capabilities. Existing methods insert tunable structures into or in parallel with the pre-trained model, which either requires back-propagation through the whole pre-trained model and is thus resource-demanding, or is limited by the temporal reasoning capability of the pre-trained structure. In this work, we present DiST, which disentangles the learning of spatial and temporal aspects of videos. Specifically, DiST uses a dual-encoder structure, where a pre-trained foundation model acts as the spatial encoder and a lightweight network is introduced as the temporal encoder. An integration branch is inserted between the encoders to fuse spatio-temporal information. The decoupled spatial and temporal learning in DiST is highly efficient because it avoids back-propagation of massive pre-trained parameters. Meanwhile, we empirically show that separated learning with an extra network for integration is beneficial to both spatial and temporal understanding. Extensive experiments on five benchmarks show that DiST delivers better performance than existing state-of-the-art methods by convincing gaps. When pre-training on the large-scale Kinetics-710, we achieve 89.7% on Kinetics-400 with a frozen ViT-L model, which verifies the scalability of DiST. Our code and models will be made available.
****
A Skeletonization Algorithm for Gradient-Based Optimization
Martin J. Menten, Johannes C. Paetzold, Veronika A. Zimmer, Suprosanna Shit, Ivan Ezhov, Robbie Holland, Monika Probst, Julia A. Schnabel, Daniel Rueckert; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21394-21403
The skeleton of a digital image is a compact representation of its topology, geometry, and scale. It has utility in many computer vision applications, such as image description, segmentation, and registration. However, skeletonization has only seen limited use in contemporary deep learning solutions. Most existing skeletonization algorithms are not differentiable, making it impossible to integrate them with gradient-based optimization. Compatible algorithms based on morphological operations and neural networks have been proposed, but their results often

deviate from the geometry and topology of the true medial axis. This work introduces the first three-dimensional skeletonization algorithm that is both compatible with gradient-based optimization and preserves an object's topology. Our method is exclusively based on matrix additions and multiplications, convolutional operations, basic non-linear functions, and sampling from a uniform probability distribution, allowing it to be easily implemented in any major deep learning library. In benchmarking experiments, we prove the advantages of our skeletonization algorithm compared to non-differentiable, morphological, and neural-network-based baselines. Finally, we demonstrate the utility of our algorithm by integrating it with two medical image processing applications that use gradient-based optimization: deep-learning-based blood vessel segmentation, and multimodal registration of the mandible in computed tomography and magnetic resonance images.
****
V3Det: Vast Vocabulary Visual Detection Dataset
Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, Dahua Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19844-19854
Recent advances in detecting arbitrary objects in the real world are trained and evaluated on object detection datasets with a relatively restricted vocabulary. To facilitate the development of more general visual object detection, we propose V3Det, a vast vocabulary visual detection dataset with precisely annotated bounding boxes on massive images. V3Det has several appealing properties: 1) Vast Vocabulary: It contains bounding boxes of objects from 13,204 categories on real-world images, which is 10 times larger than the existing large vocabulary object detection dataset, e.g., LVIS. 2) Hierarchical Category Organization: The vast vocabulary of V3Det is organized by a hierarchical category tree which annotates the inclusion relationship among categories, encouraging the exploration of category relationships in vast and open vocabulary object detection. 3) Rich Annotations: V3Det comprises precisely annotated objects in 243k images and professional descriptions of each category written by human experts and a powerful chatbot. By offering a vast exploration space, V3Det enables extensive benchmarks on both vast and open vocabulary object detection, leading to new observations, practices, and insights for future research. It has the potential to serve as a cornerstone dataset for developing more general visual perception systems. V3Det is available at https://v3det.openxlab.org.cn/.
****
Coarse-to-Fine: Learning Compact Discriminative Representation for Single-Stage Image Retrieval
Yunquan Zhu, Xinkai Gao, Bo Ke, Ruizhi Qiao, Xing Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11260-11269
Image retrieval targets to find images from a database that are visually similar to the query image. Two-stage methods following retrieve-and-rerank paradigm have achieved excellent performance, but their separate local and global modules are inefficient to real-world applications. To better trade-off retrieval efficiency and accuracy, some approaches fuse global and local feature into a joint representation to perform single-stage image retrieval. However, they are still challenging due to various situations to tackle, e.g., background, occlusion and viewpoint. In this work, we design a Coarse-to-Fine framework to learn Compact Discriminative representation (CFCD) for end-to-end single-stage image retrieval-requiring only image-level labels. Specifically, we first design a novel adaptive softmax-based loss which dynamically tunes its scale and margin within each mini-batch and increases them progressively to strengthen supervision during training and intra-class compactness. Furthermore, we propose a mechanism which attentively selects prominent local descriptors and infuse fine-grained semantic relations into the global representation by a hard negative sampling strategy to optimize inter-class distinctiveness at a global scale. Extensive experimental results have demonstrated the effectiveness of our method, which achieves state-of-the-art single-stage image retrieval performance on benchmarks such as Revisited Oxford and Revisited Paris. Code is available at https://github.com/bassyess/CFCD.
****

Multi-weather Image Restoration via Domain Translation
Prashant W. Patil, Sunil Gupta, Santu Rana, Svetha Venkatesh, Subrahmanyam Murala; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21696-21705
Weather degraded conditions such as rain, haze, snow, etc. may degrade the performance of most computer vision systems. Therefore, effective restoration of multi-weather degraded images is an essential prerequisite for successful functioning of such systems. The current multi-weather image restoration approaches utilize a model that is trained on a combined dataset consisting of individual images for rainy, snowy, and hazy weather degradations. These methods may face challenges when dealing with real-world situations where the images may have multiple, more intricate weather conditions. To address this issue, we propose a domain translation-based unified method for multi-weather image restoration. In this approach, the proposed network learns multiple weather degradations simultaneously, making it immune for real-world conditions. Specifically, we first propose an instance-level domain (weather) translation with multi-attentive feature learning approach to get different weather-degraded variants of the same scenario. Next, the original and translated images are used as input to the proposed novel multi-weather restoration network which utilizes a progressive multi-domain deformable alignment (PMDA) with cascaded multi-head attention (CMA). The proposed PMDA facilitates the restoration network to learn weather-invariant clues effectively. Further, PMDA and respective decoder features are merged via proposed CMA module for restoration. Extensive experimental results on synthetic and real-world hazy, rainy, and snowy image databases clearly demonstrate that our model outperforms the state-of-the-art multi-weather image restoration methods. The URL for our code is provided in the supplementary material and will be made public upon acceptance.
****
Deep Fusion Transformer Network with Weighted Vector-Wise Keypoints Voting for Robust 6D Object Pose Estimation
Jun Zhou, Kai Chen, Linlin Xu, Qi Dou, Jing Qin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13967-13977
One critical challenge in 6D object pose estimation from a single RGBD image is efficient integration of two different modalities, i.e., color and depth. In this work, we tackle this problem by a novel Deep Fusion Transformer (DFTr) block that can aggregate cross-modality features for improving pose estimation. Unlike existing fusion methods, the proposed DFTr can better model cross-modality semantic correlation by leveraging their semantic similarity, such that globally enhanced features from different modalities can be better integrated for improved information extraction. Moreover, to further improve robustness and efficiency, we introduce a novel weighted vector-wise voting algorithm that employs a non-iterative global optimization strategy for precise 3D keypoint localization while achieving near real-time inference. Extensive experiments show the effectiveness and strong generalization capability of our proposed 3D keypoint voting algorithm. Results on four widely used benchmarks also demonstrate that our method outperforms the state-of-the-art methods by large margins.
****
BT^2: Backward-compatible Training with Basis Transformation
Yifei Zhou, Zilu Li, Abhinav Shrivastava, Hengshuang Zhao, Antonio Torralba, Taipeng Tian, Ser-Nam Lim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11229-11238
Modern retrieval system often requires recomputing the representation of every piece of data in the gallery when updating to a better representation model. This process is known as backfilling and can be especially costly in the real world where the gallery often contains billions of samples. Recently, researchers have proposed the idea of Backward Compatible Training (BCT) where the new representation model can be trained with an auxiliary loss to make it backward compatible with the old representation. In this way, the new representation can be directly compared with the old representation, in principle avoiding the need for any backfilling. However, follow-up work shows that there is an inherent trade-off wh

ere a backward-compatible representation model cannot simultaneously maintain the performance of the new model itself. This paper reports our "not-so-surprising" finding that adding extra dimensions to the representation can help here. However, we also found that naively increasing the dimension of the representation did not work. To deal with this, we propose Backward-compatible Training with a novel Basis Transformation (BT2). A basis transformation (BT) is basically a learnable set of parameters that applies an orthonormal transformation. Such a transformation possesses an important property whereby the original information contained in its input is retained in its output. We show in this paper how a BT can be utilized to add only the necessary amount of additional dimensions. We empirically verify the advantage of BT2 over other state-of-the-art methods in a wide range of settings. We then further extend BT2 to other challenging yet more practical settings, including significant changes in model architecture (CNN to Transformers), modality change, and even a series of updates in the model architecture mimicking the evolution of deep learning models in the past decade.
****
ViperGPT: Visual Inference via Python Execution for Reasoning
Dídac Surís, Sachit Menon, Carl Vondrick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11888-11898
Answering visual queries is a complex task that requires both visual processing and reasoning. End-to-end models, the dominant approach for this task, do not explicitly differentiate between the two, limiting interpretability and generalization. Learning modular programs presents a promising alternative, but has proven challenging due to the difficulty of learning both the programs and modules simultaneously. We introduce ViperGPT, a framework that leverages code-generation models to compose vision-and-language models into subroutines to produce a result for any query. ViperGPT utilizes a provided API to access the available modules, and composes them by generating Python code that is later executed. This simple approach requires no further training, and achieves state-of-the-art results across various complex visual tasks.
****
Improving Unsupervised Visual Program Inference with Code Rewriting Families
Aditya Ganeshan, R. Kenny Jones, Daniel Ritchie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15791-15801
Programs offer compactness and structure that makes them an attractive representation for visual data. We explore how code rewriting can be used to improve systems for inferring programs from visual data. We first propose Sparse Intermittent Rewrite Injection (SIRI), a framework for unsupervised bootstrapped learning. SIRI sparsely applies code rewrite operations over a dataset of training programs, injecting the improved programs back into the training set. We design a family of rewriters for visual programming domains: parameter optimization, code pruning, and code grafting. For three shape programming languages in 2D and 3D, we experimentally validate that using SIRI with our family of rewriters improves performance: better reconstructions and faster convergence rates, compared with bootstrapped learning methods that do not use rewriters or use them naively. Finally, we demonstrate that our family of rewriters can be effectively employed at test time to improve the output of SIRI predictions. For 2D and 3D CSG, we outperform or match the reconstruction performance of recent domain-specific neural architectures, while producing more parsimonious programs, that use significantly fewer primitives.
****
Essential Matrix Estimation using Convex Relaxations in Orthogonal Space
Arman Karimian, Roberto Tron; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17142-17152
We introduce a novel method to estimate the essential matrix for two-view Structure from Motion (SfM). We show that every 3 by 3 essential matrix can be embedded in a 4 by 4 rotation, having its bottom right entry fixed to zero; we call the latter the quintessential matrix. This embedding leads to rich relations with the space of 4-D rotations, quaternions, and the classical twisted-pair ambiguity in two-view SfM. We use this structure to derive a succession of semidefinite r

elaxations that require fewer parameters than the existing non-minimal solvers a
nd yield faster convergence with certifiable optimality. We then exploit the low
-rank geometry of these relaxations to reduce them to an equivalent optimization
 on a Riemannian manifold and solve them via the Riemannian Staircase method. Th
e experimental evaluation confirms that our algorithm always finds the globally
optimal solution and outperforms the existing non-minimal methods. We make our i
mplementations open source.
****
Concept-wise Fine-tuning Matters in Preventing Negative Transfer
Yunqiao Yang, Long-Kai Huang, Ying Wei; Proceedings of the IEEE/CVF Internationa
l Conference on Computer Vision (ICCV), 2023, pp. 18753-18763
A multitude of prevalent pre-trained models mark a major milestone in the develo
pment of artificial intelligence, while fine-tuning has been a common practice t
hat enables pre-trained models to figure prominently in a wide array of target d
atasets. Our empirical results reveal that off-the-shelf fine-tuning techniques
are far from adequate to mitigate negative transfer caused by two types of under
performing features in a pre-trained model, including rare features and spurious
ly correlated features. Rooted in structural causal models of predictions after
fine-tuning, we propose a Concept-wise fine-tuning (Concept-Tuning) approach whi
ch refines feature representations in the level of patches with each patch encod
ing a concept. Concept-Tuning minimizes the negative impacts of rare features an
d spuriously correlated features by (1) maximizing the mutual information betwee
n examples in the same category with regard to a slice of rare features (a patch
) and (2) applying front-door adjustment via attention neural networks in channe
ls and feature slices (patches). The proposed Concept-Tuning consistently and si
gnificantly (by up to 4.76%) improves prior state-of-the-art fine-tuning methods
 on eleven datasets, diverse pre-training strategies (supervised and self-superv
ised ones), various network architectures, and sample sizes in a target dataset.
****
Learning Human Dynamics in Autonomous Driving Scenarios
Jingbo Wang, Ye Yuan, Zhengyi Luo, Kevin Xie, Dahua Lin, Umar Iqbal, Sanja Fidle
r, Sameh Khamis; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 20796-20806
Simulation has emerged as an indispensable tool for scaling and accelerating the
 development of self-driving systems. A critical aspect of this is simulating re
alistic and diverse human behavior and intent. In this work, we propose a holist
ic framework for learning physically plausible human dynamics from real driving
scenarios, narrowing the gap between real and simulated human behavior in safety
-critical applications. We show that state-of-the-art methods underperform in dr
iving scenarios where video data is recorded from moving vehicles, and humans ar
e frequently partially or fully occluded. Furthermore, existing methods often di
sregard the global scene where humans are situated, resulting in various motion
artifacts like foot sliding, floating, or ground penetration. Therefore, the pri
mary technical challenge of this work is to infer physically plausible human dyn
amics for the occluded body parts on uneven terrain, based on visible motions. T
o address this challenge, we propose an approach that incorporates physics with
a reinforcement learning-based motion controller to learn human dynamics for dri
ving scenarios. Our framework can simulate physically plausible human dynamics t
hat accurately match observed human motions and infill motions for occluded body
 parts, while improving the physical plausibility of the entire motion sequence.
 We evaluate our method on the challenging driving scenarios in the Waymo Open D
ataset. Experiments on the challenging Waymo Open Dataset show that our method o
utperforms state-of-the-art motion capture approaches significantly in recoverin
g high-quality, physically plausible, and scene-aware human dynamics.
****
Fine-grained Visible Watermark Removal
Li Niu, Xing Zhao, Bo Zhang, Liqing Zhang; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2023, pp. 12770-12779
Visible watermark removal aims to erase the watermark from watermarked image and
 recover the background image, which is a challenging task due to the diverse wa

termarks. Previous works have designed dynamic network to handle various types o f watermarks adaptively, but they ignore that even the watermarked region in a s ingle image can be divided into multiple local parts with distinct visual appear ances. In this work, we advance image-specific dynamic network towards part-spec ific dynamic network, which discovers multiple local parts within the watermarke d region and handle them adaptively. Specifically, we propose a query-based mult i-task framework, in which part query embeddings are jointly used in two branche s to predict part masks and restore watermarked parts. Extensive experiments dem onstrate the effectiveness of our fine-grained watermark removal network.
****
DDP: Diffusion Model for Dense Visual Prediction
Yuanfeng Ji, Zhe Chen, Enze Xie, Lanqing Hong, Xihui Liu, Zhaoqiang Liu, Tong Lu , Zhenguo Li, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21741-21752
We propose a simple, efficient, yet powerful framework for dense visual predicti ons based on the conditional diffusion pipeline. Our approach follows a "noise-t o-map" generative paradigm for prediction by progressively removing noise from a random Gaussian distribution, guided by the image. The method, called DDP, effi ciently extends the denoising diffusion process into the modern perception pipel ine. Without task-specific design and architecture customization, DDP is easy to generalize to most dense prediction tasks, e.g., semantic segmentation and dept h estimation. In addition, DDP shows attractive properties such as dynamic infer ence and uncertainty awareness, in contrast to previous single-step discriminati ve methods. We show top results on three representative tasks with six diverse b enchmarks, without tricks, DDP achieves state-of-the-art or competitive performa nce on each task compared to the specialist counterparts. For example, semantic segmentation (83.9 mIoU on Cityscapes), BEV map segmentation (70.6 mIoU on nuSce nes), and depth estimation (0.05 REL on KITTI). We hope that our approach will s erve as a solid baseline and facilitate future research.
****
Semantics-Consistent Feature Search for Self-Supervised Visual Representation Le arning
Kaiyou Song, Shan Zhang, Zimeng Luo, Tong Wang, Jin Xie; Proceedings of the IEEE /CVF International Conference on Computer Vision (ICCV), 2023, pp. 16099-16108
In contrastive self-supervised learning, the common way to learn discriminative representation is to pull different augmented "views" of the same image closer w hile pushing all other images further apart, which has been proven to be effecti ve. However, it is unavoidable to construct undesirable views containing differe nt semantic concepts during the augmentation procedure. It would damage the sema ntic consistency of representation to pull these augmentations closer in the feature space indiscriminately. In this study, we introduce feature-level a ugmentation and propose a novel semantics-consistent feature search (SCFS) metho d to mitigate this negative effect. The main idea of SCFS is to adaptively search semantics-consistent features to enhance the contrast between semantics- consistent regions in different augmentations. Thus, the trained model can learn to focus on meaningful object regions, improving the semantic representation ab ility. Extensive experiments conducted on different datasets and tasks demonstra te that SCFS effectively improves the performance of self-supervised learning an d achieves state-of-the-art performance on different downstream tasks.
****
GridMM: Grid Memory Map for Vision-and-Language Navigation
Zihan Wang, Xiangyang Li, Jiahao Yang, Yeqi Liu, Shuqiang Jiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15625 -15636
Vision-and-language navigation (VLN) enables the agent to navigate to a remote l ocation following the natural language instruction in 3D environments. To repres ent the previously visited environment, most approaches for VLN implement memory using recurrent states, topological maps, or top-down semantic maps. In contras t to these approaches, we build the top-down egocentric and dynamically growing Grid Memory Map (i.e., GridMM) to structure the visited environment. From a glob

al perspective, historical observations are projected into a unified grid map in a top-down view, which can better represent the spatial relations of the enviro nment. From a local perspective, we further propose an instruction relevance agg regation method to capture fine-grained visual clues in each grid region. Extens ive experiments are conducted on both the REVERIE, R2R, SOON datasets in the dis crete environments, and the R2R-CE dataset in the continuous environments, showi ng the superiority of our proposed method.
****
Probabilistic Modeling of Inter- and Intra-observer Variability in Medical Image Segmentation
Arne Schmidt, Pablo Morales-Álvarez, Rafael Molina; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21097-21106
Medical image segmentation is a challenging task, particularly due to inter- and intra-observer variability, even between medical experts. In this paper, we pro pose a novel model, called Probabilistic Inter-Observer and iNtra-Observer varia tion NetwOrk (Pionono). It captures the labeling behavior of each rater with a m ultidimensional probability distribution and integrates this information with th e feature maps of the image to produce probabilistic segmentation predictions. T he model is optimized by variational inference and can be trained end-to-end. It outperforms state-of-the-art models such as STAPLE, Probabilistic U-Net, and mo dels based on confusion matrices. Additionally, Pionono predicts multiple cohere nt segmentation maps that mimic the rater's expert opinion, which provides addit ional valuable information for the diagnostic process. Experiments on real-world cancer segmentation datasets demonstrate the high accuracy and efficiency of Pi onono, making it a powerful tool for medical image analysis.
****
LAC - Latent Action Composition for Skeleton-based Action Segmentation
Di Yang, Yaohui Wang, Antitza Dantcheva, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, Francois Bremond; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2023, pp. 13679-13690
Skeleton-based action segmentation requires recognizing composable actions in un trimmed videos. Current approaches decouple this problem by first extracting loc al visual features from skeleton sequences and then processing them by a tempora l model to classify frame-wise actions. However, their performances remain limit ed as the visual features cannot sufficiently express composable actions. In thi s context, we propose Latent Action Composition (LAC), a novel self-supervised f ramework aiming at learning from synthesized composable motions for skeleton-bas ed action segmentation. LAC is composed of a novel generation module towards syn thesizing new sequences. Specifically, we design a linear latent space in the ge nerator to represent primitive motion. New composed motions can be synthesized b y simply performing arithmetic operations on latent representations of multiple input skeleton sequences. LAC leverages such synthesized sequences, which have l arge diversity and complexity, for learning visual representations of skeletons in both sequence and frame spaces via contrastive learning. The resulting visual encoder has a high expressive power and can be effectively transferred onto act ion segmentation tasks by end-to-end fine-tuning without the need for additional temporal models. We conduct a study focusing on transfer-learning and we show t hat representations learned from pre-trained LAC outperform the state-of-the-art by a large margin on TSU, Charades, PKU-MMD datasets.
****
Learning Vision-and-Language Navigation from YouTube Videos
Kunyang Lin, Peihao Chen, Diwei Huang, Thomas H. Li, Mingkui Tan, Chuang Gan; Pr oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20 23, pp. 8317-8326
Vision-and-language navigation (VLN) requires an embodied agent to navigate in r ealistic 3D environments using natural language instructions. Existing VLN metho ds suffer from training on small-scale environments or unreasonable path-instruc tion datasets, limiting the generalization to unseen environments. There are mas sive house tour videos on YouTube, providing abundant real navigation experience s and layout information. However, these videos have not been explored for VLN b

efore. In this paper, we propose to learn an agent from these videos by creating a large-scale dataset which comprises reasonable path-instruction pairs from house tour videos and pre-training the agent on it. To achieve this, we have to tackle the challenges of automatically constructing path-instruction pairs and exploiting real layout knowledge from raw and unlabeled videos. To address these, we first leverage an entropy-based method to construct the nodes of a path trajectory. Then, we propose an action-aware generator for generating instructions from unlabeled trajectories. Last, we devise a trajectory judgment pretext task to encourage the agent to mine the layout knowledge. Experimental results show that our method achieves state-of-the-art performance on two popular benchmarks (R2R and REVERIE). Code is available at https://github.com/JeremyLinky/YouTube-VLN.
****
Total-Recon: Deformable Scene Reconstruction for Embodied View Synthesis
Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, Deva Ramanan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17671-17682
We explore the task of embodied view synthesis from monocular videos of deformable scenes. Given a minute-long RGBD video of people interacting with their pets, we render the scene from novel camera trajectories derived from the in-scene motion of actors: (1) egocentric cameras that simulate the point of view of a target actor and (2) 3rd-person cameras that follow the actor. Building such a system requires reconstructing the root-body and articulated motion of every actor, as well as a scene representation that supports free-viewpoint synthesis. Longer videos are more likely to capture the scene from diverse viewpoints (which helps reconstruction) but are also more likely to contain larger motions (which complicates reconstruction). To address these challenges, we present Total-Recon, the first method to photorealistically reconstruct deformable scenes from long monocular RGBD videos. Crucially, to scale to long videos, our method hierarchically decomposes the scene into the background and objects, whose motion is decomposed into carefully initialized root-body motion and local articulations. To quantify such "in-the-wild" reconstruction and view synthesis, we collect ground-truth data from a specialized stereo RGBD capture rig for 11 challenging videos, significantly outperforming prior methods.
****
AdaNIC: Towards Practical Neural Image Compression via Dynamic Transform Routing
Lvfang Tao, Wei Gao, Ge Li, Chenhao Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16879-16888
Compressive autoencoders (CAEs) play an important role in deep learning-based image compression, but large-scale CAEs are computationally expensive. We propose a framework with three techniques to enable efficient CAE-based image coding: 1) Spatially-adaptive convolution and normalization operators enable block-wise nonlinear transform to spend FLOPs unevenly across the image to be compressed, according to a transform capacity map. 2) Just-unpenalized model capacity (JUMC) optimizes the transform capacity of each CAE block via rate-distortion-complexity optimization, finding the optimal capacity for the source image content. 3) A lightweight routing agent model predicts the transform capacity map for the CAEs by approximating JUMC targets. By activating the best-sized sub-CAE inside the slimmable supernet, our approach achieves up to 40% computational speed-up with minimal BD-Rate increase, validating its ability to save computational resources in a content-aware manner.
****
Uncertainty-aware State Space Transformer for Egocentric 3D Hand Trajectory Forecasting
Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, Yu Kong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13702-13711
Hand trajectory forecasting from egocentric views is crucial for enabling a prompt understanding of human intentions when interacting with AR/VR systems. However, existing methods handle this problem in a 2D image space which is inadequate for 3D real-world applications. In this paper, we set up an egocentric 3D hand t

rajectory forecasting task that aims to predict hand trajectories in a 3D space from early observed RGB videos in a first-person view. To fulfill this goal, we propose an uncertainty-aware state space Transformer (USST) that takes the merits of the attention mechanism and aleatoric uncertainty within the framework of the classical state-space model. The model can be further enhanced by the velocity constraint and visual prompt tuning (VPT) on large vision transformers. Moreover, we develop an annotation workflow to collect 3D hand trajectories with high quality. Experimental results on H2O and EgoPAT3D datasets demonstrate the superiority of USST for both 2D and 3D trajectory forecasting. The code and datasets are publicly released: https://actionlab-cv.github.io/EgoHandTrajPred.
****
Pretrained Language Models as Visual Planners for Human Assistance
Dhruvesh Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, Ruta Desai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15302-15314
In our pursuit of advancing multi-modal AI assistants capable of guiding users to achieve complex multi-step goals, we propose the task of 'Visual Planning for Assistance (VPA)'. Given a succinct natural language goal, e.g., "make a shelf", and a video of the user's progress so far, the aim of VPA is to devise a plan, i.e., a sequence of actions such as "sand shelf", "paint shelf", etc. to realize the specified goal. This requires assessing the user's progress from the (untrimmed) video, and relating it to the requirements of natural language goal, i.e., which actions to select and in what order? Consequently, this requires handling long video history and arbitrarily complex action dependencies. To address these challenges, we decompose VPA into video action segmentation and forecasting. Importantly, we experiment by formulating the forecasting step as a multi-modal sequence modeling problem, allowing us to leverage the strength of pre-trained LMs (as the sequence model). This novel approach, which we call Visual Language Model based Planner (VLaMP), outperforms baselines across a suite of metrics that gauge the quality of the generated plans. Furthermore, through comprehensive ablations, we also isolate the value of each component--language pre-training, visual observations, and goal information. We have open-sourced all the data, model checkpoints, and training code.
****
Dynamic Point Fields
Sergey Prokudin, Qianli Ma, Maxime Raafat, Julien Valentin, Siyu Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7964-7976
Recent years have witnessed significant progress in the field of neural surface reconstruction. While extensive focus was put on volumetric and implicit approaches, a number of works have shown that explicit graphics primitives, such as point clouds, can significantly reduce computational complexity without sacrificing the reconstructed surface quality. However, less emphasis has been put on modeling dynamic surfaces with point primitives. In this work, we present a dynamic point field model that combines the representational benefits of explicit point-based graphics with implicit deformation networks to allow efficient modeling of non-rigid 3D surfaces. Using explicit surface primitives also allows us to easily incorporate well-established constraints such as isometric-as-possible regularization. While learning this deformation model is prone to local optima when trained in a fully unsupervised manner, we propose to also leverage semantic information, such as keypoint correspondence, to guide the deformation learning. We demonstrate how this approach can be used for creating an expressive animatable human avatar from a collection of 3D scans. Here, previous methods mostly rely on variants of the linear blend skinning paradigm, which fundamentally limits the expressivity of such models when dealing with complex cloth appearances, such as long skirts. We show the advantages of our dynamic point field framework in terms of its representational power, learning efficiency, and robustness to out-of-distribution novel poses. The code for the project is publicly available.
****
Lip2Vec: Efficient and Robust Visual Speech Recognition via Latent-to-Latent Vis

ual to Audio Representation Mapping
Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Haithem Boussaid, Ebtessam Almazrouei, Merouane Debbah; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13790-13801
Visual Speech Recognition (VSR) differs from the common perception tasks as it requires deeper reasoning over the video sequence, even by human experts. Despite the recent advances in VSR, current approaches rely on labeled data to fully train or finetune their models predicting the target speech. This hinders their ability to generalize well beyond the training set and leads to performance degeneration under out-of-distribution challenging scenarios. Un-like previous works that involve auxiliary losses or com-plex training procedures and architectures, we propose a simple approach, named Lip2Vec that is based on learning a prior model. Given a robust visual speech encoder, this network maps the encoded latent representations of the lip sequence to their corresponding latents from the audio pair, which are sufficiently invariant for effective text decoding. The generated audio representation is then decoded to text using an off-the-shelf Audio Speech Recognition (ASR) model. The proposed model compares favorably with fully-supervised learning methods on the LRS3 dataset achieving 26 WER. Unlike SoTA approaches, our model keeps a rea-sonable performance on the VoxCeleb2-en test set. We believe that reprogramming the VSR as an ASR task narrows the performance gap between the two and paves the way for more flexible formulations of lip reading.
****
Privacy Preserving Localization via Coordinate Permutations
Linfei Pan, Johannes L. Schönberger, Viktor Larsson, Marc Pollefeys; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18174-18183
Recent methods on privacy-preserving image-based localization use a random line parameterization to protect the privacy of query images and database maps. The lifting of points to lines effectively drops one of the two geometric constraints traditionally used with point-to-point correspondences in structure-based localization. This leads to a significant loss of accuracy for the privacy-preserving methods. In this paper, we overcome this limitation by devising a coordinate permutation scheme that allows for recovering the original point positions during pose estimation. The recovered points provide the full 2D geometric constraints and enable us to close the gap between privacy-preserving and traditional methods in terms of accuracy. Another limitation of random line methods is their vulnerability to density based 3D line cloud inversion attacks. Our method not only provides better accuracy than the original random line based approach but also provides stronger privacy guarantees against these recently proposed attacks. Extensive experiments on standard benchmark datasets demonstrate these improvements consistently across both scenarios of protecting the privacy of query images as well as the database map.
****
Random Boxes Are Open-world Object Detectors
Yanghao Wang, Zhongqi Yue, Xian-Sheng Hua, Hanwang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6233-6243
We show that classifiers trained with random region proposals achieve state-of-the-art Open-world Object Detection (OWOD): they can not only maintain the accuracy of the known objects (w/ training labels), but also considerably improve the recall of unknown ones (w/o training labels). Specifically, we propose RandBox, a Fast R-CNN based architecture trained on random proposals at each training iteration, surpassing existing Faster R-CNN and Transformer based OWOD. Its effectiveness stems from the following two benefits introduced by randomness. First, as the randomization is independent of the distribution of the limited known objects, the random proposals become the instrumental variable that prevents the training from being confounded by the known objects. Second, the unbiased training encourages more proposal explorations by using our proposed matching score that does not penalize the random proposals whose prediction scores do not match the known objects. On two benchmarks: Pascal-VOC/MS-COCO and LVIS, RandBox significan

tly outperforms the previous state-of-the-art in all metrics. We also detail the ablations on randomization and loss designs. Codes and other details are in Appendix.
****
DiffDreamer: Towards Consistent Unsupervised Single-view Scene Extrapolation with Conditional Diffusion Models
Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, Gordon Wetzstein; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2139-2150
Scene extrapolation---the idea of generating novel views by flying into a given image---is a promising, yet challenging task. For each predicted frame, a joint inpainting and 3D refinement problem has to be solved, which is ill posed and includes a high level of ambiguity. Moreover, training data for long-range scenes is difficult to obtain and usually lacks sufficient views to infer accurate camera poses. We introduce DiffDreamer, an unsupervised framework capable of synthesizing novel views depicting a long camera trajectory while training solely on internet-collected images of nature scenes. Utilizing the stochastic nature of the guided denoising steps, we train the diffusion models to refine projected RGBD images but condition the denoising steps on multiple past and future frames for inference. We demonstrate that image-conditioned diffusion models can effectively perform long-range scene extrapolation while preserving consistency significantly better than prior GAN-based methods. DiffDreamer is a powerful and efficient solution for scene extrapolation, producing impressive results despite limited supervision. Project page: https://primecai.github.io/diffdreamer.
****
Spectral Graphormer: Spectral Graph-Based Transformer for Egocentric Two-Hand Reconstruction using Multi-View Color Images
Tze Ho Elden Tse, Franziska Mueller, Zhengyang Shen, Danhang Tang, Thabo Beeler, Mingsong Dou, Yinda Zhang, Sasa Petrovic, Hyung Jin Chang, Jonathan Taylor, Bardia Doosti; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14666-14677
We propose a novel transformer-based framework that reconstructs two high fidelity hands from multi-view RGB images. Unlike existing hand pose estimation methods, where one typically trains a deep network to regress hand model parameters from single RGB image, we consider a more challenging problem setting where we directly regress the absolute root poses of two-hands with extended forearm at high resolution from egocentric view. As existing datasets are either infeasible for egocentric viewpoints or lack background variations, we create a large-scale synthetic dataset with diverse scenarios and collect a real dataset from multi-calibrated camera setup to verify our proposed multi-view image feature fusion strategy. To make the reconstruction physically plausible, we propose two strategies: (i) a coarse-to-fine spectral graph convolution decoder to smoothen the meshes during upsampling and (ii) an optimisation-based refinement stage at inference to prevent self-penetrations. Through extensive quantitative and qualitative evaluations, we show that our framework is able to produce realistic two-hand reconstructions and demonstrate the generalisation of synthetic-trained models to real data, as well as real-time AR/VR applications.
****
SMMix: Self-Motivated Image Mixing for Vision Transformers
Mengzhao Chen, Mingbao Lin, Zhihang Lin, Yuxin Zhang, Fei Chao, Rongrong Ji; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17260-17270
CutMix is a vital augmentation strategy that determines the performance and generalization ability of vision transformers (ViTs). However, the inconsistency between the mixed images and the corresponding labels harms its efficacy. Existing CutMix variants tackle this problem by generating more consistent mixed images or more precise mixed labels, but inevitably introduce heavy training overhead or require extra information, undermining ease of use. To this end, we propose an efficient and effective Self-Motivated image Mixing method (SMMix), which motivates both image and label enhancement by the model under training itself. Specifi

cally, we propose a max-min attention region mixing approach that enriches the a ttention-focused objects in the mixed images. Then, we introduce a fine-grained label assignment technique that co-trains the output tokens of mixed images with fine-grained supervision. Moreover, we devise a novel feature consistency const raint to align features from mixed and unmixed images. Due to the subtle designs of the self-motivated paradigm, our SMMix is significant in its smaller trainin g overhead and better performance than other CutMix variants. In particular, SMM ix improves the accuracy of DeiT-T/S/B, CaiT-XXS-24/36, and PVT-T/S/M/L by more than +1% on ImageNet-1k. The generalization capability of our method is also dem onstrated on downstream tasks and out-of-distribution datasets. Our project is a vailable at https://github.com/ChenMnZ/SMMix.
****

Enhancing Adversarial Robustness in Low-Label Regime via Adaptively Weighted Reg ularization and Knowledge Distillation
Dongyoon Yang, Insung Kong, Yongdai Kim; Proceedings of the IEEE/CVF Internation al Conference on Computer Vision (ICCV), 2023, pp. 4552-4561
Adversarial robustness is a research area that has recently received a lot of at tention in the quest for trustworthy artificial intelligence. However, recent wo rks on adversarial robustness have focused on supervised learning where it is as sumed that labeled data is plentiful. In this paper, we investigate semi -superv ised adversarial training where labeled data is scarce. We derive two upper boun ds for the robust risk and propose a regularization term for unlabeled data moti vated by these two upper bounds. Then, we develop a semi-supervised adversarial training algorithm that combines the proposed regularization term with knowledge distillation using a semi-supervised teacher. Our experiments show that our pro posed algorithm achieves state-of-the-art performance with significant margins c ompared to existing algorithms. In particular, compared to supervised learning a lgorithms, performance of our proposed algorithm is not much worse even when the amount of labeled data is very small. For example, our algorithm with only 8% l abeled data is comparable to supervised adversarial training algorithms that use all labeled data, both in terms of standard and robust accuracies on CIFAR-10.
****

Recovering a Molecule's 3D Dynamics from Liquid-phase Electron Microscopy Movies
Enze Ye, Yuhang Wang, Hong Zhang, Yiqin Gao, Huan Wang, He Sun; Proceedings of t he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10767- 10777
The dynamics of biomolecules are crucial for our understanding of their function ing in living systems. However, current 3D imaging techniques, such as cryogenic electron microscopy (cryo-EM), require freezing the sample, which limits the ob servation of their conformational changes in real time. The innovative liquid-ph ase electron microscopy (liquid-phase EM) technique allows molecules to be place d in the native liquid environment, providing a unique opportunity to observe th eir dynamics. In this paper, we propose TEMPOR, a Temporal Electron MicroscoPy O bject Reconstruction algorithm for liquid-phase EM that leverages an implicit ne ural representation (INR) and a dynamical variational auto-encoder (DVAE) to rec over time series of molecular structures. We demonstrate its advantages in recov ering different motion dynamics from two simulated datasets, 7bcq and Cas9. To o ur knowledge, our work is the first attempt to directly recover 3D structures of a temporally-varying particle from liquid-phase EM movies. It provides a promis ing new approach for studying molecules' 3D dynamics in structural biology.
****

Reconciling Object-Level and Global-Level Objectives for Long-Tail Detection
Shaoyu Zhang, Chen Chen, Silong Peng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18982-18992
Large vocabulary object detectors are often faced with the long-tailed label dis tributions, seriously degrading their ability to detect rarely seen categories. On one hand, the rare objects are prone to be misclassified as frequent categori es. On the other hand, due to the limitation on the total number of detections p er image, detectors usually rank all the confidence scores globally and filter o ut the lower-ranking ones. This may result in missed detection during inference,

 especially for the rare categories that naturally come with lower scores. Exist
ing methods mainly focus on the former problem and design various classification
 loss to enhance the object-level classification accuracy, but largely overlook
the global-level ranking task. In this paper, we propose a novel framework that
Reconciles Object-level and Global-level (ROG) objectives to address both proble
ms. As a multi-task learning framework, ROG simultaneously trains the model with
 two tasks: classifying each object proposal individually and ranking all the co
nfidence scores globally. Specifically, complementary to the object-level classi
fication loss for model discrimination, we design a generalized average precisio
n (GAP) loss to explicitly optimize the global-level score ranking across differ
ent objects. For each category, GAP loss generates balanced gradients to rectify
 the ranking errors. In experiments, we show that GAP loss is highly versatile t
o be plugged into various advanced methods and brings considerable benefits.
****
In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video
Retrieval
Nina Shvetsova, Anna Kukleva, Bernt Schiele, Hilde Kuehne; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21981-21992
Large-scale noisy web image-text datasets have been proven to be efficient for l
earning robust vision-language models. However, to transfer them to the task of
video retrieval, models still need to be fine-tuned on hand-curated paired text-
video data to adapt to the diverse styles of video descriptions. To address this
 problem without the need for hand-annotated pairs, we propose a new setting, te
xt-video retrieval with uncurated & unpaired data, that uses only text queries t
ogether with uncurated web videos during training without any paired text-video
data. To this end, we propose an approach, In-Style, that learns the style of th
e text queries and transfers it to uncurated web videos. Moreover, to improve ge
neralization, we show that one model can be trained with multiple text styles. T
o this end, we introduce a multi-style contrastive training procedure, that impr
oves the generalizability over several datasets simultaneously. We evaluate our
model on retrieval performance over multiple datasets to demonstrate the advanta
ges of our style transfer framework on the new task of uncurated & unpaired text
-video retrieval and improve state-of-the-art performance on zero-shot text-vide
o retrieval.
****
MIMO-NeRF: Fast Neural Rendering with Multi-input Multi-output Neural Radiance F
ields
Takuhiro Kaneko; Proceedings of the IEEE/CVF International Conference on Compute
r Vision (ICCV), 2023, pp. 3273-3283
Neural radiance fields (NeRFs) have shown impressive results for novel view synt
hesis. However, they depend on the repetitive use of a single-input single-outpu
t multilayer perceptron (SISO MLP) that maps 3D coordinates and view direction t
o the color and volume density in a sample-wise manner, which slows the renderin
g. We propose a multi-input multi-output NeRF (MIMO-NeRF) that reduces the numbe
r of MLPs running by replacing the SISO MLP with a MIMO MLP and conducting mappi
ngs in a group-wise manner. One notable challenge with this approach is that the
 color and volume density of each point can differ according to a choice of inpu
t coordinates in a group, which can lead to some notable ambiguity. We also prop
ose a self-supervised learning method that regularizes the MIMO MLP with multipl
e fast reformulated MLPs to alleviate this ambiguity without using pretrained mo
dels. The results of a comprehensive experimental evaluation including comparati
ve and ablation studies are presented to show that MIMO-NeRF obtains a good trad
e-off between speed and quality with a reasonable training time. We then demonst
rate that MIMO-NeRF is compatible with and complementary to previous advancement
s in NeRFs by applying it to two representative fast NeRFs, i.e., a NeRF with a
sampling network (DONeRF) and a NeRF with alternative representations (TensoRF).
****
Instance Neural Radiance Field
Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, Chi-Keung Tang; Proceedings of
 the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 787-

This paper presents one of the first learning-based NeRF 3D instance segmentation pipelines, dubbed as Instance Neural Radiance Field, or Instance-NeRF. Taking a NeRF pretrained from multi-view RGB images as input, Instance-NeRF can learn 3D instance segmentation of a given scene, represented as an instance field component of the NeRF model. To this end, we adopt a 3D proposal-based mask prediction network on the sampled volumetric features from NeRF, which generates discrete 3D instance masks. The coarse 3D mask prediction is then projected to image space to match 2D segmentation masks from different views generated by existing panoptic segmentation models, which are used to supervise the training of the instance field. Notably, beyond generating consistent 2D segmentation maps from novel views, Instance-NeRF can query instance information at any 3D point, which greatly enhances NeRF object segmentation and manipulation. Our method is also one of the first to achieve such results in pure inference. Experimented on synthetic and real-world NeRF datasets with complex indoor scenes, Instance-NeRF surpasses previous NeRF segmentation works and competitive 2D segmentation methods in segmentation performance on unseen views. Code and data are available at https://github.com/lyclyc52/Instance_NeRF.
****
One-bit Flip is All You Need: When Bit-flip Attack Meets Model Training
Jianshuo Dong, Han Qiu, Yiming Li, Tianwei Zhang, Yuanjie Li, Zeqi Lai, Chao Zhang, Shu-Tao Xia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4688-4698
Deep neural networks (DNNs) are widely deployed on real-world devices. Concerns regarding their security have gained great attention from researchers. Recently, a new weight modification attack called bit flip attack (BFA) was proposed, which exploits memory fault inject techniques such as row hammer to attack quantized models in the deployment stage. With only a few bit flips, the target model can be rendered useless as a random guesser or even be implanted with malicious functionalities. In this work, we seek to further reduce the number of bit flips. We propose a training-assisted bit flip attack, in which the adversary is involved in the training stage to build a high-risk model to release. This high-risk model, obtained coupled with a corresponding malicious model, behaves normally and can escape various detection methods. The results on benchmark datasets show that an adversary can easily convert this high-risk but normal model to a malicious one on victim's side by flipping only one critical bit on average in the deployment stage. Moreover, our attack still poses a significant threat even when defenses are employed. The codes for reproducing main experiments are available at https://github.com/jianshuod/TBA.
****
CLIPTER: Looking at the Bigger Picture in Scene Text Recognition
Aviad Aberdam, David Bensaid, Alona Golts, Roy Ganz, Oren Nuriel, Royee Tichauer, Shai Mazor, Ron Litman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21706-21717
Reading text in real-world scenarios often requires understanding the context surrounding it, especially when dealing with poor-quality text. However, current scene text recognizers are unaware of the bigger picture as they operate on cropped text images. In this study, we harness the representative capabilities of modern vision-language models, such as CLIP, to provide scene-level information to the crop-based recognizer. We achieve this by fusing a rich representation of the entire image, obtained from the vision-language model, with the recognizer word-level features via a gated cross-attention mechanism. This component gradually shifts to the context-enhanced representation, allowing for stable fine-tuning of a pretrained recognizer. We demonstrate the effectiveness of our model-agnostic framework, CLIPTER (CLIP TExt Recognition), on leading text recognition architectures and achieve state-of-the-art results across multiple benchmarks. Furthermore, our analysis highlights improved robustness to out-of-vocabulary words and enhanced generalization in low-data regimes.
****
Revisiting Scene Text Recognition: A Data Perspective

Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, Lianwen Jin; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20543-
20554
This paper aims to re-assess scene text recognition (STR) from a data-oriented p
erspective. We begin by revisiting the six commonly used benchmarks in STR and o
bserve a trend of performance saturation, whereby only 2.91% of the benchmark im
ages cannot be accurately recognized by an ensemble of 13 representative models.
 While these results are impressive and suggest that STR could be considered sol
ved, however, we argue that this is primarily due to the less challenging nature
 of the common benchmarks, thus concealing the underlying issues that STR faces.
 To this end, we consolidate a large-scale real STR dataset, namely Union14M, wh
ich comprises 4 million labeled images and 10 million unlabeled images, to asses
s the performance of STR models in more complex real-world scenarios. Our experi
ments demonstrate that the 13 models can only achieve an average accuracy of 66.
53% on the 4 million labeled images, indicating that STR still faces numerous ch
allenges in the real world. By analyzing the error patterns of the 13 models, we
 identify seven open challenges in STR and develop a challenge-driven benchmark
consisting of eight distinct subsets to facilitate further progress in the field
. Our exploration demonstrates that STR is far from being solved and leveraging
data may be a promising solution. In this regard, we find that utilizing the 10
million unlabeled images through self-supervised pre-training can significantly
improve the robustness of STR model in real-world scenarios and leads to state-o
f-the-art performance. Code and dataset is available at https: //github.com/Moun
tchicken/Union14M .
****
Improving CLIP Fine-tuning Performance
Yixuan Wei, Han Hu, Zhenda Xie, Ze Liu, Zheng Zhang, Yue Cao, Jianmin Bao, Dong
Chen, Baining Guo; Proceedings of the IEEE/CVF International Conference on Compu
ter Vision (ICCV), 2023, pp. 5439-5449
CLIP models have demonstrated impressively high zero-shot recognition accuracy,
however, their fine-tuning performance on downstream vision tasks is sub-optimal
. Contrarily, masked image modeling (MIM) performs exceptionally for fine-tuning
 on downstream tasks, despite the absence of semantic labels during training. We
 note that the two tasks have different ingredients: image-level targets versus
token-level targets, a cross-entropy loss versus a regression loss, and full-ima
ge inputs versus partial-image inputs. To mitigate the differences, we introduce
 a classical feature map distillation framework, which can simultaneously inheri
t the semantic capability of CLIP models while constructing a task incorporated
key ingredients of MIM. Experiments suggest that the feature map distillation ap
proach significantly boosts the fine-tuning performance of CLIP models on severa
l typical downstream vision tasks. We also observe that the approach yields new
CLIP representations which share some diagnostic properties with those of MIM. F
urthermore, the feature map distillation approach generalizes to other pre-train
ing models, such as DINO, DeiT and SwinV2-G, reaching a new record of 64.2 mAP o
n COCO object detection with +1.1 improvement. The code and mod- els are publicl
y available at https://github.com/ SwinTransformer/Feature-Distillation.
****
The Power of Sound (TPoS): Audio Reactive Video Generation with Stable Diffusion
Yujin Jeong, Wonjeong Ryoo, Seunghyun Lee, Dabin Seo, Wonmin Byeon, Sangpil Kim,
 Jinkyu Kim; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2023, pp. 7822-7832
In recent years, video generation has become a prominent generative tool and has
 drawn significant attention. However, there is little consideration in audio-to
-video generation, though audio contains unique qualities like temporal semantic
s and magnitude. Hence, we propose The Power of Sound (TPoS) model to incorporat
e audio input that includes both changeable temporal semantics and magnitude. To
 generate video frames, TPoS utilizes a latent stable diffusion model with textu
al semantic information, which is then guided by the sequential audio embedding
from our pretrained Audio Encoder. As a result, this method produces audio react
ive video contents. We demonstrate the effectiveness of TPoS across various task

s and compare its results with current state-of-the-art techniques in the field of audio-to-video generation. More examples are available at https://ku-vai.github.io/TPoS/

****

SOCS: Semantically-Aware Object Coordinate Space for Category-Level 6D Object Pose Estimation under Large Shape Variations

Boyan Wan, Yifei Shi, Kai Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14065-14074

Most learning-based approaches to category-level 6D pose estimation are design around normalized object coordinate space (NOCS). While being successful, NOCS-based methods become inaccurate and less robust when handling objects of a category containing significant intra-category shape variations. This is because the object coordinates induced by global and rigid alignment of objects are semantically incoherent, making the coordinate regression hard to learn and generalize. We propose Semantically-aware Object Coordinate Space (SOCS) built by warping-and-aligning the objects guided by a sparse set of keypoints with semantically meaningful correspondence. SOCS is semantically coherent: Any point on the surface of a object can be mapped to a semantically meaningful location in SOCS, allowing for accurate pose and size estimation under large shape variations. To learn effective coordinate regression to SOCS, we propose a novel multi-scale coordinate-based attention network. Evaluations demonstrate that our method is easy to train, well-generalizing for large intra-category shape variations and robust to inter-object occlusions.

****

NeRF-LOAM: Neural Implicit Representation for Large-Scale Incremental LiDAR Odometry and Mapping

Junyuan Deng, Qi Wu, Xieyuanli Chen, Songpengcheng Xia, Zhen Sun, Guoqing Liu, Wenxian Yu, Ling Pei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8218-8227

Simultaneously odometry and mapping using LiDAR data is an important task for mobile systems to achieve full autonomy in large-scale environments. However, most existing LiDAR-based methods prioritize tracking quality over reconstruction quality. Although the recently developed neural radiance fields (NeRF) have shown promising advances in implicit reconstruction for indoor environments, the problem of simultaneous odometry and mapping for large-scale scenarios using incremental LiDAR data remains unexplored. To bridge this gap, in this paper, we propose a novel NeRF-based LiDAR odometry and mapping approach, NeRF-LOAM, consisting of three modules neural odometry, neural mapping, and mesh reconstruction. All these modules utilize our proposed neural signed distance function, which separates LiDAR points into ground and non-ground points to reduce Z-axis drift, optimizes odometry and voxel embeddings concurrently, and in the end generates dense smooth mesh maps of the environment. Moreover, this joint optimization allows our NeRF-LOAM to be pre-trained free and exhibit strong generalization abilities when applied to different environments. Extensive evaluations on three publicly available datasets demonstrate that our approach achieves state-of-the-art odometry and mapping performance, as well as a strong generalization in large-scale environments utilizing LiDAR data. Furthermore, we perform multiple ablation studies to validate the effectiveness of our network design. The implementation of our approach will be made available at https://github.com/JunyuanDeng/NeRF-LOAM.

****

DINAR: Diffusion Inpainting of Neural Textures for One-Shot Human Avatars

David Svitov, Dmitrii Gudkov, Renat Bashirov, Victor Lempitsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7062-7072

We present DINAR, an approach for creating realistic rigged fullbody avatars from single RGB images. Similarly to previous works, our method uses neural textures combined with the SMPL-X body model to achieve photo-realistic quality of avatars while keeping them easy to animate and fast to infer. To restore the texture, we use a latent diffusion model and show how such model can be trained in the neural texture space. The use of the diffusion model allows us to realistically

reconstruct large unseen regions such as the back of a person given the frontal view. The models in our pipeline are trained using 2D images and videos only. In the experiments, our approach achieves state-of-the-art rendering quality and good generalization to new poses and viewpoints. In particular, the approach improves state-of-the-art on the SnapshotPeople public benchmark.
****
DPM-OT: A New Diffusion Probabilistic Model Based on Optimal Transport
Zezeng Li, Shenghao Li, Zhanpeng Wang, Na Lei, Zhongxuan Luo, David Xianfeng Gu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22624-22633
Sampling from diffusion probabilistic models (DPMs) can be viewed as a piecewise distribution transformation, which generally requires hundreds or thousands of steps of the inverse diffusion trajectory to get a high-quality image. Recent progress in designing fast samplers for DPMs achieves a trade-off between sampling speed and sample quality by knowledge distillation or adjusting the variance schedule or the denoising equation. However, it can't be optimal in both aspects and often suffer from mode mixture in short steps. To tackle this problem, we innovatively regard inverse diffusion as an optimal transport (OT) problem between latents at different stages and propose DPM-OT, a unified learning framework for fast DPMs with the direct expressway represented by OT map, which can generate high-quality samples within around 10 function evaluations. By calculating the semi-discrete optimal transport between the data latents and the white noise, we obtain the expressway from the prior distribution to the data distribution, while significantly alleviating the problem of mode mixture. In addition, we give the error bound of the proposed method, which theoretically guarantees the stability of the algorithm. Extensive experiments validate the effectiveness and advantages of DPM-OT in terms of speed and quality (FID and mode mixture), thus representing an efficient solution for generative modeling. Source codes are available at https://github.com/cognaclee/DPM-OT
****
ElasticViT: Conflict-aware Supernet Training for Deploying Fast Vision Transformer on Diverse Mobile Devices
Chen Tang, Li Lyna Zhang, Huiqiang Jiang, Jiahang Xu, Ting Cao, Quanlu Zhang, Yuqing Yang, Zhi Wang, Mao Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5829-5840
Neural Architecture Search (NAS) has shown promising performance in the automatic design of vision transformers (ViT) exceeding 1G FLOPs. However, designing lightweight and low-latency ViT models for diverse mobile devices remains a big challenge. In this work, we propose ElasticViT, a two-stage NAS approach that trains a high-quality ViT supernet over a very large search space for covering a wide range of mobile devices, and then searches an optimal sub-network (subnet) for direct deployment. However, current supernet training methods that rely on uniform sampling suffer from the gradient conflict issue: the sampled subnets can have vastly different model sizes (e.g., 50M vs. 2G FLOPs), leading to different optimization directions and inferior performance. To address this challenge, we propose two novel sampling techniques: complexity-aware sampling and performance-aware sampling. Complexity-aware sampling limits the FLOPs difference among the subnets sampled across adjacent training steps, while covering different-sized subnets in the search space. Performance-aware sampling further selects subnets that have good accuracy, which can reduce gradient conflicts and improve supernet quality. Our discovered models, ElasticViT models, achieve top-1 accuracy from 67.2% to 80.0% on ImageNet from 60M to 800M FLOPs without extra retraining, outperforming all prior CNNs and ViTs in terms of accuracy and latency. Our tiny and small models are also the first ViT models that surpass state-of-the-art CNNs with significantly lower latency on mobile devices. For instance, ElasticViT-S1 runs 2.62x faster than EfficientNet-B0 with 0.1% higher accuracy.
****
OmniLabel: A Challenging Benchmark for Language-Based Object Detection
Samuel Schulter, Vijay Kumar B G, Yumin Suh, Konstantinos M. Dafnis, Zhixing Zhang, Shiyu Zhao, Dimitris Metaxas; Proceedings of the IEEE/CVF International Conf

erence on Computer Vision (ICCV), 2023, pp. 11953-11962
Language-based object detection is a promising direction towards building a natural interface to describe objects in images that goes far beyond plain category names. While recent methods show great progress in that direction, proper evaluation is lacking. With OmniLabel, we propose a novel task definition, dataset, and evaluation metric. The task subsumes standard and open-vocabulary detection as well as referring expressions. With more than 30K unique object descriptions on over 25K images, OmniLabel provides a challenge benchmark with diverse and complex object descriptions in a naturally open-vocabulary setting. Moreover, a key differentiation to existing benchmarks is that our object descriptions can refer to one, multiple or even no object, hence, providing negative examples in free-form text. The proposed evaluation handles the large label space and judges performance via a modified average precision metric, which we validate by evaluating strong language-based baselines. OmniLabel indeed provides a challenging test bed for future research on language-based detection.
****
Noise-Aware Learning from Web-Crawled Image-Text Data for Image Captioning
Wooyoung Kang, Jonghwan Mun, Sungjun Lee, Byungseok Roh; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2942-2952
Image captioning is one of the straightforward tasks that can take advantage of large-scale web-crawled data which provides rich knowledge about the visual world for a captioning model. However, since web-crawled data contains image-text pairs that are aligned at different levels, the inherent noises (e.g., misaligned pairs) make it difficult to learn a precise captioning model. While the filtering strategy can effectively remove noisy data, it leads to a decrease in learnable knowledge and sometimes brings about a new problem of data deficiency.
To take the best of both worlds, we propose a Noise-aware Captioning (NoC) framework, which learns rich knowledge from the whole web-crawled data while being less affected by the noises. This is achieved by the proposed alignment-level-controllable captioner, which is learned using alignment levels of the image-text pairs as a control signal during training. The alignment-level-conditioned training allows the model to generate high-quality captions by simply setting the control signal to the desired alignment level at inference time.
An in-depth analysis shows the effectiveness of our framework in handling noise.
With two tasks of zero-shot captioning and text-to-image retrieval using generated captions (i.e., self-retrieval), we also demonstrate our model can produce high-quality captions in terms of descriptiveness and distinctiveness. The code is available at https://github.com/kakaobrain/noc.
****
Divide&Classify: Fine-Grained Classification for City-Wide Visual Geo-Localization
Gabriele Trivigno, Gabriele Berton, Juan Aragon, Barbara Caputo, Carlo Masone; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11142-11152
Visual Place recognition is commonly addressed as an image retrieval problem. However, retrieval methods are impractical to scale to large datasets, densely sampled from city-wide maps, since their dimension impact negatively on the inference time. Using approximate nearest neighbour search for retrieval helps to mitigate this issue, at the cost of a performance drop.
In this paper we investigate whether we can effectively approach this task as a classification problem, thus bypassing the need for a similarity search. We find that existing classification methods for coarse, planet-wide localization are not suitable for the fine-grained and city-wide setting. This is largely due to how the dataset is split into classes, because these methods are designed to handle a sparse distribution of photos and as such do not consider the visual aliasing problem across neighbouring classes that naturally arises in dense scenarios. Thus, we propose a partitioning scheme that enables a fast and accurate inference, preserving a simple learning procedure, and a novel inference pipeline based on an ensemble of novel classifiers that uses the prototypes learned via an an

gular margin loss. Our method, Divide&Classify (D&C), enjoys the fast inference of classification solutions and an accuracy competitive with retrieval methods on the fine-grained, city-wide setting.
 Moreover, we show that D&C can be paired with existing retrieval pipelines to speed up computations by over 20 times while increasing their recall, leading to new state-of-the-art results. Code is available at https://github.com/ga1i13o/Divide-and-Classify
****
3D Semantic Subspace Traverser: Empowering 3D Generative Model with Shape Editing Capability
Ruowei Wang, Yu Liu, Pei Su, Jianwei Zhang, Qijun Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14406-14417
Shape generation is the practice of producing 3D shapes as various representations for 3D content creation. Previous studies on 3D shape generation have focused on shape quality and structure, without or less considering the importance of semantic information. Consequently, such generative models often fail to preserve the semantic consistency of shape structure or enable manipulation of the semantic attributes of shapes during generation. In this paper, we proposed a novel semantic generative model named 3D Semantic Subspace Traverser that utilizes semantic attributes for category-specific 3D shape generation and editing. Our method utilizes implicit functions as the 3D shape representation and combines a novel latent-space GAN with a linear subspace model to discover semantic dimensions in the local latent space of 3D shapes. Each dimension of the subspace corresponds to a particular semantic attribute, and we can edit the attributes of generated shapes by traversing the coefficients of those dimensions. Experimental results demonstrate that our method can produce plausible shapes with complex structures and enable the editing of semantic attributes. The code and trained models are available at https://github.com/TrepangCat/3D_Semantic_Subspace_Traverser.
****
Inherent Redundancy in Spiking Neural Networks
Man Yao, Jiakui Hu, Guangshe Zhao, Yaoyuan Wang, Ziyang Zhang, Bo Xu, Guoqi Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16924-16934
Spiking Neural Networks (SNNs) are well known as a promising energy-efficient alternative to conventional artificial neural networks. Subject to the preconceived impression that SNNs are sparse firing, the analysis and optimization of inherent redundancy in SNNs have been largely overlooked, thus the potential advantages of spike-based neuromorphic computing in accuracy and energy efficiency are interfered. In this work, we pose and focus on three key questions regarding the inherent redundancy in SNNs. We argue that the redundancy is induced by the spatio-temporal invariance of SNNs, which enhances the efficiency of parameter utilization but also invites lots of noise spikes. Further, we analyze the effect of spatio-temporal invariance on the spatio-temporal dynamics and spike firing of SNNs. Then, motivated by these analyses, we propose an Advance Spatial Attention (ASA) module to harness SNNs' redundancy, which can adaptively optimize their membrane potential distribution by a pair of individual spatial attention sub-modules. In this way, noise spike features are accurately regulated. Experimental results demonstrate that the proposed method can significantly drop the spike firing with better performance than state-of-the-art baselines. Our code is available in https://github.com/BICLab/ASA-SNN.
****
Text2Room: Extracting Textured 3D Meshes from 2D Text-to-Image Models
Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, Matthias Nießner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7909-7920
We present Text2Room, a method for generating room-scale textured 3D meshes from a given text prompt as input. To this end, we leverage pre-trained 2D text-to-image models to synthesize a sequence of images from different poses. In order to lift these outputs into a consistent 3D scene representation, we combine monocular depth estimation with a text-conditioned inpainting model. The core idea of

our approach is a tailored viewpoint selection such that the content of each image can be fused into a seamless, textured 3D mesh. More specifically, we propose a continuous alignment strategy that iteratively fuses scene frames with the existing geometry to create a seamless mesh. Unlike existing works that focus on generating single objects [56, 41] or zoom-out trajectories [18] from text, our method generates complete 3D scenes with multiple objects and explicit 3D geometry. We evaluate our approach using qualitative and quantitative metrics, demonstrating it as the first method to generate room-scale 3D geometry with compelling textures from only text as input.
****

On the Robustness of Normalizing Flows for Inverse Problems in Imaging

Seongmin Hong, Inbum Park, Se Young Chun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10745-10755

Conditional normalizing flows can generate diverse image samples for solving inverse problems. Most normalizing flows for inverse problems in imaging employ the conditional affine coupling layer that can generate diverse images quickly. However, unintended severe artifacts are occasionally observed in the output of them. In this work, we address this critical issue by investigating the origins of these artifacts and proposing the conditions to avoid them. First of all, we empirically and theoretically reveal that these problems are caused by "exploding inverse" in the conditional affine coupling layer for certain out-of-distribution (OOD) conditional inputs. Then, we further validated that the probability of causing erroneous artifacts in pixels is highly correlated with a Mahalanobis distance-based OOD score for inverse problems in imaging. Lastly, based on our investigations, we propose a remark to avoid exploding inverse and then based on it, we suggest a simple remedy that substitutes the affine coupling layers with the modified rational quadratic spline coupling layers in normalizing flows, to encourage the robustness of generated image samples. Our experimental results demonstrated that our suggested methods effectively suppressed critical artifacts occurring in normalizing flows for super-resolution space generation and low-light image enhancement.
****

FastRecon: Few-shot Industrial Anomaly Detection via Fast Feature Reconstruction

Zheng Fang, Xiaoyang Wang, Haocheng Li, Jiejie Liu, Qiugui Hu, Jimin Xiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17481-17490

In industrial anomaly detection, data efficiency and the ability for fast migration across products become the main concerns when developing detection algorithms. Existing methods tend to be data-hungry and work in the one-model-one-category way, which hinders their effectiveness in real-world industrial scenarios. In this paper, we propose a few-shot anomaly detection strategy that works in a low-data regime and can generalize across products at no cost. Given a defective query sample, we propose to utilize a few normal samples as a reference to reconstruct its normal version, where the final anomaly detection can be achieved by sample alignment. Specifically, we introduce a novel regression with distribution regularization to obtain the optimal transformation from support to query features, which guarantees the reconstruction result shares visual similarity with the query sample and meanwhile maintains the property of normal samples. Experimental results reflect that our method significantly outperforms previous state-of-the-art at both image and pixel-level AUROC performances from 2 to 8-shot scenarios. Besides, with only a limited number of training samples (less than 8 samples), our method reaches competitive performance with vanilla AD methods which are trained with extensive normal samples.
****

Local or Global: Selective Knowledge Assimilation for Federated Learning with Limited Labels

Yae Jee Cho, Gauri Joshi, Dimitrios Dimitriadis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17087-17096

Many existing FL methods assume clients with fully-labeled data, while in realistic settings, clients have limited labels due to the expensive and laborious pro

cess of labeling. Limited labeled local data of the clients often leads to their
 local model having poor generalization abilities to their larger unlabeled loca
l data, such as having class-distribution mismatch with the unlabeled data. As a
 result, clients may instead look to benefit from the global model trained acros
s clients to leverage their unlabeled data, but this also becomes difficult due
to data heterogeneity across clients. In our work, we propose FedLabel where cli
ents selectively choose the local or global model to pseudo-label their unlabele
d data depending on which is more of an expert of the data. We further utilize b
oth the local and global models' knowledge via global-local consistency regulari
zation which minimizes the divergence between the two models' outputs when they
have identical pseudo-labels for the unlabeled data. Unlike other semi-supervise
d FL baselines, our method does not require additional experts other than the lo
cal or global model, nor require additional parameters to be communicated. We al
so do not assume any server-labeled data or fully labeled clients. For both cros
s-device and cross-silo settings, we show that FedLabel outperforms other semi-s
upervised FL baselines by 8-24%, and even outperforms standard fully supervised
FL baselines (100% labeled data) with only 5-20% of labeled data.
****
DistillBEV: Boosting Multi-Camera 3D Object Detection with Cross-Modal Knowledge
 Distillation
Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, Xiaodong Yang; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8637-864
6
3D perception based on the representations learned from multi-camera bird's-eye-
view (BEV) is trending as cameras are cost-effective for mass production in auto
nomous driving industry. However, there exists a distinct performance gap betwee
n multi-camera BEV and LiDAR based 3D object detection. One key reason is that L
iDAR captures accurate depth and other geometry measurements, while it is notori
ously challenging to infer such 3D information from merely image input. In this
work, we propose to boost the representation learning of a multi-camera BEV base
d student detector by training it to imitate the features of a well-trained LiDA
R based teacher detector. We propose effective balancing strategy to enforce the
 student to focus on learning the crucial features from the teacher, and general
ize knowledge transfer to multi-scale layers with temporal fusion. We conduct ex
tensive evaluations on multiple representative models of multi-camera BEV. Exper
iments reveal that our approach renders significant improvement over the student
 models, leading to the state-of-the-art performance on the popular benchmark nu
Scenes.
****
PoseFix: Correcting 3D Human Poses with Natural Language
Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, Grégory Rogez; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023
, pp. 15018-15028
Automatically producing instructions to modify one's posture could open the door
 to endless applications, such as personalized coaching and in-home physical the
rapy. Tackling the reverse problem (i.e., refining a 3D pose based on some natur
al language feedback) could help for assisted 3D character animation or robot te
aching, for instance.
 Although a few recent works explore the connections between natural language an
d 3D human pose, none focus on describing 3D body pose differences. In this pape
r, we tackle the problem of correcting 3D human poses with natural language.
 To this end, we introduce the PoseFix dataset, which consists of several thousa
nd paired 3D poses and their corresponding text feedback, that describe how the
source pose needs to be modified to obtain the target pose. We demonstrate the p
otential of this dataset on two tasks: (1) text-based pose editing, that aims at
 generating corrected 3D body poses given a query pose and a text modifier; and
(2) correctional text generation, where instructions are generated based on the
differences between two body poses. The dataset and the code are available at ht
tps://europe.naverlabs.com/research/computer-vision/posefix/.
****

TAPIR: Tracking Any Point with Per-Frame Initialization and Temporal Refinement
Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, Andrew Zisserman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10061-10072
We present a novel model for Tracking Any Point (TAP) that effectively tracks any queried point on any physical surface throughout a video sequence. Our approach employs two stages: (1) a matching stage, which independently locates a suitable candidate point match for the query point on every other frame, and (2) a refinement stage, which updates both the trajectory and query features based on local correlations. The resulting model surpasses all baseline methods by a significant margin on the TAP-Vid benchmark, as demonstrated by an approximate 20% absolute average Jaccard (AJ) improvement on DAVIS. Our model facilitates fast inference on long and high-resolution video sequences. On a modern GPU, our implementation has the capacity to track points faster than real-time. Given the high-quality trajectories extracted from a large dataset, we demonstrate a proof-of-concept diffusion model which generates trajectories from static images, enabling plausible animations. Visualizations, source code, and pretrained models can be found at https://deepmind-tapir.github.io.
****
SwinLSTM: Improving Spatiotemporal Prediction Accuracy using Swin Transformer and LSTM
Song Tang, Chuang Li, Pu Zhang, RongNian Tang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13470-13479
Integrating CNNs and RNNs to capture spatiotemporal dependencies is a prevalent strategy for spatiotemporal prediction tasks. However, the property of CNNs to learn local spatial information decreases their efficiency in capturing spatiotemporal dependencies, thereby limiting their prediction accuracy. In this paper, we propose a new recurrent cell, SwinLSTM, which integrates Swin Transformer blocks and the simplified LSTM, an extension that replaces the convolutional structure in ConvLSTM with the self-attention mechanism. Furthermore, we construct a network with SwinLSTM cell as the core for spatiotemporal prediction. Without using unique tricks, SwinLSTM outperforms state-of-the-art methods on Moving MNIST, Human3.6m, TaxiBJ, and KTH datasets. In particular, it exhibits a significant improvement in prediction accuracy compared to ConvLSTM. Our competitive experimental results demonstrate that learning global spatial dependencies is more advantageous for models to capture spatiotemporal dependencies. We hope that SwinLSTM can serve as a solid baseline to promote the advancement of spatiotemporal prediction accuracy. The codes are publicly available at https://github.com/SongTang-x/SwinLSTM.
****
Detecting Objects with Context-Likelihood Graphs and Graph Refinement
Aritra Bhowmik, Yu Wang, Nora Baka, Martin R. Oswald, Cees G. M. Snoek; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6524-6533
The goal of this paper is to detect objects by exploiting their interrelationships. Contrary to existing methods, which learn objects and relations separately, our key idea is to learn the object-relation distribution jointly. We first propose a novel way of creating a graphical representation of an image from inter-object relation priors and initial class predictions, we call a context-likelihood graph. We then learn the joint distribution with an energy-based modeling technique which allows to sample and refine the context-likelihood graph iteratively for a given image. Our formulation of jointly learning the distribution enables us to generate a more accurate graph representation of an image which leads to a better object detection performance. We demonstrate the benefits of our context-likelihood graph formulation and the energy-based graph refinement via experiments on the Visual Genome and MS-COCO datasets where we achieve a consistent improvement over object detectors like DETR and Faster-RCNN, as well as alternative methods modeling object interrelationships separately. Our method is detector agnostic, end-to-end trainable, and especially beneficial for rare object classes.
****

Coarse-to-Fine Amodal Segmentation with Shape Prior
Jianxiong Gao, Xuelin Qian, Yikai Wang, Tianjun Xiao, Tong He, Zheng Zhang, Yanwei Fu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1262-1271
Amodal object segmentation is a challenging task that involves segmenting both visible and occluded parts of an object.
In this paper, we propose a novel approach, called Coarse-to-Fine Segmentation (C2F-Seg), that addresses this problem by progressively modeling the amodal segmentation.
C2F-Seg initially reduces the learning space from the pixel-level image space to the vector-quantized latent space.
This enables us to better handle long-range dependencies and learn a coarse-grained amodal segment from visual features and visible segments.
However, this latent space lacks detailed information about the object, which makes it difficult to provide a precise segmentation directly.
To address this issue, we propose a convolution refine module to inject fine-grained information and provide a more precise amodal object segmentation based on visual features and coarse-predicted segmentation.
To help the studies of amodal object segmentation, we create a synthetic amodal dataset, named as MOViD-Amodal (MOViD-A), which can be used for both image and video amodal object segmentation.
We extensively evaluate our model on two benchmark datasets: KINS and COCO-A. Our empirical results demonstrate the superiority of C2F-Seg.
Moreover, we exhibit the potential of our approach for video amodal object segmentation tasks on FISHBOWL and our proposed MOViD-A.
Project page at: https://jianxgao.github.io/C2F-Seg.
****
DEDRIFT: Robust Similarity Search under Content Drift
Dmitry Baranchuk, Matthijs Douze, Yash Upadhyay, I. Zeki Yalniz; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11026-11035
The statistical distribution of content uploaded and searched on media sharing sites changes over time due to seasonal, sociological and technical factors. We investigate the impact of this "content drift" for large-scale similarity search tools, based on nearest neighbor search in embedding space. Unless a costly index reconstruction is performed frequently, content drift degrades the search accuracy and efficiency. The degradation is especially severe since, in general, both the query and database distributions change.

We introduce and analyze real-world image and video datasets for which temporal information is available over a long time period. Based on the learnings, we devise DeDrift, a method that updates embedding quantizers to continuously adapt large-scale indexing structures on-the-fly. DeDrift almost eliminates the accuracy degradation due to the query and database content drift while being up to 100x faster than a full index reconstruction.
****
Learning Pseudo-Relations for Cross-domain Semantic Segmentation
Dong Zhao, Shuang Wang, Qi Zang, Dou Quan, Xiutiao Ye, Rui Yang, Licheng Jiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19191-19203
Domain adaptive semantic segmentation aims to adapt a model trained on labeled source domain to the unlabeled target domain. Self-training shows competitive potential in this field. Existing methods along this stream mainly focus on selecting reliable predictions on target data as pseudo-labels for category learning, while ignoring the useful relations between pixels for relation learning. In this paper, we propose a pseudo-relation learning framework, Relation Teacher (RTea), which can exploitable pixel relations to efficiently use unreliable pixels and learn generalized representations. In this framework, we build reasonable pseudo-relations on local grids and fuse them with low-level relations in the image space, which are motivated by the reliable local relations prior and available lo

w-level relations prior. Then, we design a pseudo-relation learning strategy and optimize the class probability to meet the relation consistency by finding the optimal sub-graph division. In this way, the model's certainty and consistency of prediction are
enhanced on the target domain, and the cross-domain inadaptation is further eliminated. Extensive experiments on three datasets demonstrate the effectiveness of the proposed method.
****
AdVerb: Visually Guided Audio Dereverberation
Sanjoy Chowdhury, Sreyan Ghosh, Subhrajyoti Dasgupta, Anton Ratnarajah, Utkarsh Tyagi, Dinesh Manocha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7884-7896
We present AdVerb, a novel audio-visual dereverberation framework that uses visual cues in addition to the reverberant sound to estimate clean audio. Although audio-only dereverberation is a well-studied problem, our approach incorporates the complementary visual modality to perform audio dereverberation. Given an image of the environment where the reverberated sound signal has been recorded, AdVerb employs a novel geometry-aware cross-modal transformer architecture that captures scene geometry and audio-visual cross-modal relationship to generate a complex ideal ratio mask, which, when applied to the reverberant audio predicts the clean sound. The effectiveness of our method is demonstrated through extensive quantitative and qualitative evaluations. Our approach significantly outperforms traditional audio-only and audio-visual baselines on three downstream tasks: speech enhancement, speech recognition, and speaker verification, with relative improvements in the range of 18% - 82% on the LibriSpeech test-clean set. We also achieve highly satisfactory RT60 error scores on the AVSpeech dataset.
****
Audio-Enhanced Text-to-Video Retrieval using Text-Conditioned Feature Alignment
Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, Mohamed Omar; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12054-12064
Text-to-video retrieval systems have recently made significant progress by utilizing pre-trained models trained on large-scale image-text pairs. However, most of the latest methods primarily focus on the video modality while disregarding the audio signal for this task. Nevertheless, a recent advancement by EclipSE has improved long-range text-to-video retrieval by developing an audiovisual video representation. Nonetheless, the objective of the text-to-video retrieval task is to capture the complementary audio and video information that is pertinent to the text query rather than simply achieving better audio and video alignment. To address this issue, we introduce TEFAL, a TExt-conditioned Feature ALignment method that produces both audio and video representations conditioned on the text query. Instead of using only an audiovisual attention block, which could suppress the audio information relevant to the text query, our approach employs two independent cross-modal attention blocks that enable the text to attend to the audio and video representations separately. Our proposed method's efficacy is demonstrated on four benchmark datasets that include audio: MSR-VTT, LSMDC, VATEX, and Charades, and achieves better than state-of-the-art performance consistently across the four datasets. This is attributed to the additional text-query-conditioned audio representation and the complementary information it adds to the text-query-conditioned video representation.
****
Open-vocabulary Object Segmentation with Diffusion Models
Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, Weidi Xie; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7667-7676
The goal of this paper is to extract the visual-language correspondence from a pre-trained text-to-image diffusion model, in the form of segmentation map, i.e., simultaneously generating images and segmentation masks for the corresponding visual entities described in the text prompt. We make the following contributions: (i) we pair the existing Stable Diffusion model with a novel grounding module,

that can be trained to align the visual and textual embedding space of the diffusion model with only a small number of object categories; (ii) we establish an automatic pipeline for constructing a dataset, that consists of image, segmentation mask, text prompt triplets, to train the proposed grounding module; (iii) we evaluate the performance of open-vocabulary grounding on images generated from the text-to-image diffusion model and show that the module can well segment the objects of categories beyond seen ones at training time; (iv) we adopt the augmented diffusion model to build a synthetic semantic segmentation dataset, and show that, training a standard segmentation model on such dataset demonstrates competitive performance on the zero-shot segmentation (ZS3) benchmark, which opens up new opportunities for adopting the powerful diffusion model for discriminative tasks.
****
Human-centric Scene Understanding for 3D Large-scale Scenarios
Yiteng Xu, Peishan Cong, Yichen Yao, Runnan Chen, Yuenan Hou, Xinge Zhu, Xuming He, Jingyi Yu, Yuexin Ma; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20349-20359
Human-centric scene understanding is significant for real-world applications, but it is extremely challenging due to the existence of diverse human poses and actions, complex human-environment interactions, severe occlusions in crowds, etc. In this paper, we present a large-scale multi-modal dataset for human-centric scene understanding, dubbed HuCenLife, which is collected in diverse daily-life scenarios with rich and fine-grained annotations. Our HuCenLife can benefit many 3D perception tasks, such as segmentation, detection, action recognition, etc., and we also provide benchmarks for these tasks to facilitate related research. In addition, we design novel modules for LiDAR-based segmentation and action recognition, which are more applicable for large-scale human-centric scenarios and achieve state-of-the-art performance. The dataset and code can be found at https://github.com/4DVLab/HuCenLife.git.
****
With a Little Help from Your Own Past: Prototypical Memory Networks for Image Captioning
Manuele Barraco, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Rita Cucchiara; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3021-3031
Image captioning, like many tasks involving vision and language, currently relies on Transformer-based architectures for extracting the semantics in an image and translating it into linguistically coherent descriptions. Although successful, the attention operator only considers a weighted summation of projections of the current input sample, therefore ignoring the relevant semantic information which can come from the joint observation of other samples. In this paper, we devise a network which can perform attention over activations obtained while processing other training samples, through a prototypical memory model. Our memory models the distribution of past keys and values through the definition of prototype vectors which are both discriminative and compact. Experimentally, we assess the performance of the proposed model on the COCO dataset, in comparison with carefully designed baselines and state-of-the-art approaches, and by investigating the role of each of the proposed components. We demonstrate that our proposal can increase the performance of an encoder-decoder Transformer by 3.7 CIDEr points both when training in cross-entropy only and when fine-tuning with self-critical sequence training. Source code and trained models are available at: https://github.com/aimagelab/PMA-Net.
****
SimMatchV2: Semi-Supervised Learning with Graph Consistency
Mingkai Zheng, Shan You, Lang Huang, Chen Luo, Fei Wang, Chen Qian, Chang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16432-16442
Semi-Supervised image classification is one of the most fundamental problem in computer vision, which significantly reduces the need for human labor. In this paper, we introduce a new semi-supervised learning algorithm - SimMatchV2, which f

ormulates various consistency regularizations between labeled and unlabeled data from the graph perspective. In SimMatchV2, we regard the augmented view of a sample as a node, which consists of a label and its corresponding representation. Different nodes are connected with the edges, which are measured by the similarity of the node representations. Inspired by the message passing and node classification in graph theory, we propose four types of consistencies, namely 1) node-node consistency, 2) node-edge consistency, 3) edge-edge consistency, and 4) edge-node consistency. We also uncover that a simple feature normalization can reduce the gaps of the feature norm between different augmented views, significantly improving the performance of SimMatchV2. Our SimMatchV2 has been validated on multiple semi-supervised learning benchmarks. Notably, with ResNet-50 as our backbone and 300 epochs of training, SimMatchV2 achieves 71.9% and 76.2% Top-1 Accuracy with 1% and 10% labeled examples on ImageNet, which significantly outperforms the previous methods and achieves state-of-the-art performance.
****

Reinforced Disentanglement for Face Swapping without Skip Connection

Xiaohang Ren, Xingyu Chen, Pengfei Yao, Heung-Yeung Shum, Baoyuan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20665-20675

The SOTA face swap models still suffer the problem of either target identity (i.e., shape) being leaked or the target non-identity attributes (i.e., background, hair) failing to be fully preserved in the final results. We show that this insufficient disentanglement is caused by two flawed designs that were commonly adopted in prior models: (1) counting on only one compressed encoder to represent both the semantic-level non-identity facial attributes(i.e., pose) and the pixel-level non-facial region details, which is contradictory to satisfy at the same time; (2) highly relying on long skip-connections between the encoder and the final generator, leaking a certain amount of target face identity into the result. To fix them, we introduce a new face swap framework called "WSC-swap" that gets rid of skip connections and uses two target encoders to respectively capture the pixel-level non-facial region attributes and the semantic non-identity attributes in the face region. To further reinforce the disentanglement learning for the target encoder, we employ both identity removal loss via adversarial training (i.e., GAN) and the non-identity preservation loss via prior 3DMM models like. Extensive experiments on both FaceForensics++ and CelebA-HQ show that our results significantly outperform previous works on a rich set of metrics, including one novel metric for measuring identity consistency that was completely neglected before.
****

PDiscoNet: Semantically consistent part discovery for fine-grained recognition

Robert van der Klis, Stephan Alaniz, Massimiliano Mancini, Cassio F. Dantas, Dino Ienco, Zeynep Akata, Diego Marcos; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1866-1876

Fine-grained classification often requires recognizing specific object parts, such as beak shape and wing patterns for birds. Encouraging a fine-grained classification model to first detect such parts and then using them to infer the class could help us gauge whether the model is indeed looking at the right details better than with interpretability methods that provide a single attribution map.

We propose PDiscoNet to discover object parts by using only image-level class labels along with priors encouraging the parts to be: discriminative, compact, distinct from each other, equivariant to rigid transforms, and active in at least some of the images. In addition to using the appropriate losses to encode these priors, we propose to use part-dropout, where full part feature vectors are dropped at once to prevent a single part from dominating in the classification, and part feature vector modulation, which makes the information coming from each part distinct from the perspective of the classifier.

Our results on CUB, CelebA, and PartImageNet show that the proposed method provides substantially better part discovery performance than previous methods while not requiring any additional hyper-parameter tuning and without penalizing the classification performance.

****
Privacy-Preserving Face Recognition Using Random Frequency Components
Yuxi Mi, Yuge Huang, Jiazhen Ji, Minyi Zhao, Jiaxiang Wu, Xingkun Xu, Shouhong Ding, Shuigeng Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19673-19684
The ubiquitous use of face recognition has sparked increasing privacy concerns, as unauthorized access to sensitive face images could compromise the information of individuals. This paper presents an in-depth study of the privacy protection of face images' visual information and against recovery. Drawing on the perceptual disparity between humans and models, we propose to conceal visual information by pruning human-perceivable low-frequency components. For impeding recovery, we first elucidate the seeming paradox between reducing model-exploitable information and retaining high recognition accuracy. Based on recent theoretical insights and our observation on model attention, we propose a solution to the dilemma, by advocating for the training and inference of recognition models on randomly selected frequency components. We distill our findings into a novel privacy-preserving face recognition method, PartialFace. Extensive experiments demonstrate that PartialFace effectively balances privacy protection goals and recognition accuracy. Code is available at: https://github.com/Tencent/TFace.
****
Vision Transformer Adapters for Generalizable Multitask Learning
Deblina Bhattacharjee, Sabine Süsstrunk, Mathieu Salzmann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19015-19026
We introduce the first multitasking vision transformer adapters that learn generalizable task affinities which can be applied to novel tasks and domains. Integrated into an off-the-shelf vision transformer backbone, our adapters can simultaneously solve multiple dense vision tasks in a parameter-efficient manner, unlike existing multitasking transformers that are parametrically expensive. In contrast to concurrent methods, we do not require retraining or fine-tuning whenever a new task or domain is added. We introduce a task-adapted attention mechanism within our adapter framework that combines gradient-based task similarities with attention-based ones. The learned task affinities generalize to the following settings: zero-shot task transfer, unsupervised domain adaptation, and generalization without fine-tuning to novel domains. We demonstrate that our approach outperforms not only the existing convolutional neural network-based multitasking methods but also the vision transformer-based ones. Our project page is at https://ivrl.github.io/VTAGML.
****
How to Choose your Best Allies for a Transferable Attack?
Thibault Maho, Seyed-Mohsen Moosavi-Dezfooli, Teddy Furon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4542-4551
The transferability of adversarial examples is a key issue in the security of deep neural networks. The possibility of an adversarial example crafted for a source model fooling another targeted model makes the threat of adversarial attacks more realistic. Measuring transferability is a crucial problem, but the Attack Success Rate alone does not provide a sound evaluation. This paper proposes a new methodology for evaluating transferability by putting distortion in a central position. This new tool shows that transferable attacks may perform far worse than a black box attack if the attacker randomly picks the source model. To address this issue, we propose a new selection mechanism, called FiT, which aims at choosing the best source model with only a few preliminary queries to the target. Our experimental results show that FiT is highly effective at selecting the best source model for multiple scenarios such as single-model attacks, ensemble-model attacks and multiple attacks.
****
CVRecon: Rethinking 3D Geometric Feature Learning For Neural Reconstruction
Ziyue Feng, Liang Yang, Pengsheng Guo, Bing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17750-17760
Recent advances in neural reconstruction using posed image sequences have made remarkable progress. However, due to the lack of depth information, existing volu

metric-based techniques simply duplicate 2D image features of the object surface along the entire camera ray. We contend this duplication introduces noise in empty and occluded spaces, posing challenges for producing high-quality 3D geometry. Drawing inspiration from traditional multi-view stereo methods, we propose an end-to-end 3D neural reconstruction framework CVRecon, designed to exploit the rich geometric embedding in the cost volumes to facilitate 3D geometric feature learning. Furthermore, we present Ray-contextual Compensated Cost Volume (RCCV), a novel 3D geometric feature representation that encodes view-dependent information with improved integrity and robustness. Through comprehensive experiments, we demonstrate that our approach significantly improves the reconstruction quality in various metrics and recovers clear fine details of the 3D geometries. Our extensive ablation studies provide insights into the development of effective 3D geometric feature learning schemes. Project page: https://cvrecon.ziyue.cool
****
Self-Supervised Object Detection from Egocentric Videos
Peri Akiva, Jing Huang, Kevin J Liang, Rama Kovvuri, Xingyu Chen, Matt Feiszli, Kristin Dana, Tal Hassner; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5225-5237
Understanding the visual world from the perspective of humans (egocentric) has been a long-standing challenge in computer vision. Egocentric videos exhibit high scene complexity and irregular motion flows compared to typical video understanding tasks. With the egocentric domain in mind, we address the problem of self-supervised, class-agnostic object detection, which aims to locate all objects in a given view, regardless of category, without any annotations or pre-training weights. Our method, self-supervised object Detection from Egocentric VIdeos (DEVI), generalizes appearance-based methods to learn features that are category-specific and invariant to viewing angles and illumination conditions from highly ambiguous environments in an end-to-end manner. Our approach leverages typical human behavior and its egocentric perception to sample diverse views of the same objects for our multi-view and scale-regression loss functions. With our learned cluster residual module, we are able to effectively describe multi-category patches for better complex scene understanding. DEVI provides a boost in performance on recent egocentric datasets, with performance gains up to 4.11% AP50, 0.11% AR1, 1.32% AR10, and 5.03% AR100, while significantly reducing model complexity. We also demonstrate competitive performance on out-of-domain datasets without additional training or fine-tuning.
****
Prior-guided Source-free Domain Adaptation for Human Pose Estimation
Dripta S. Raychaudhuri, Calvin-Khang Ta, Arindam Dutta, Rohit Lal, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14996-15006
Domain adaptation methods for 2D human pose estimation typically require continuous access to the source data during adaptation, which can be challenging due to privacy, memory, or computational constraints. To address this limitation, we focus on the task of source-free domain adaptation for pose estimation, where a source model must adapt to a new target domain using only unlabeled target data. Although recent advances have introduced source-free methods for classification tasks, extending them to the regression task of pose estimation is non-trivial. In this paper, we present Prior-guided Self-training (POST), a pseudo-labeling approach that builds on the popular Mean Teacher framework to compensate for the distribution shift. POST leverages prediction-level and feature-level consistency between a student and teacher model against certain image transformations. In the absence of source data, POST utilizes a human pose prior that regularizes the adaptation process by directing the model to generate more accurate and anatomically plausible pose pseudo-labels. Despite being simple and intuitive, our framework can deliver significant performance gains compared to applying the source model directly to the target data, as demonstrated in our extensive experiments and ablation studies. In fact, our approach achieves comparable performance to recent state-of-the-art
methods that use source data for adaptation

****
ClothesNet: An Information-Rich 3D Garment Model Repository with Simulated Clothes Environment
Bingyang Zhou, Haoyu Zhou, Tianhai Liang, Qiaojun Yu, Siheng Zhao, Yuwei Zeng, Jun Lv, Siyuan Luo, Qiancai Wang, Xinyuan Yu, Haonan Chen, Cewu Lu, Lin Shao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20428-20438
We present ClothesNet: a large-scale dataset of 3D clothes objects with information-rich annotations. Our dataset consists of around 4000 models covering 11 categories annotated with clothes features, boundary lines, and keypoints. ClothesNet can be used to facilitate a variety of computer vision and robot interaction tasks. Using our dataset, we establish benchmark tasks for clothes perception, including classification, boundary line segmentation, and keypoint detection, and develop simulated clothes environments for robotic interaction tasks, including rearranging, folding, hanging, and dressing.
We also demonstrate the efficacy of our ClothesNet in real-world experiments.
****
Auxiliary Tasks Benefit 3D Skeleton-based Human Motion Prediction
Chenxin Xu, Robby T. Tan, Yuhong Tan, Siheng Chen, Xinchao Wang, Yanfeng Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9509-9520
Exploring spatial-temporal dependencies from observed motions is one of the core challenges of human motion prediction. Previous methods mainly focus on dedicated network structures to model the spatial and temporal dependencies. This paper considers a new direction by introducing a model learning framework with auxiliary tasks. In our auxiliary tasks, partial body joints' coordinates are corrupted by either masking or adding noise and the goal is to recover corrupted coordinates depending on the rest coordinates. To work with auxiliary tasks, we propose a novel auxiliary-adapted transformer, which can handle incomplete, corrupted motion data and achieve coordinate recovery via capturing spatial-temporal dependencies. Through auxiliary tasks, the auxiliary-adapted transformer is promoted to capture more comprehensive spatial-temporal dependencies among body joints' coordinates, leading to better feature learning. Extensive experimental results have shown that our method outperforms state-of-the-art methods by remarkable margins of 7.2%, 3.7%, and 9.4% in terms of 3D mean per joint position error (MPJPE) on the Human3.6M, CMU Mocap, and 3DPW datasets, respectively. We also demonstrate that our method is more robust under data missing cases and noisy data cases. Code is available at https://github.com/MediaBrain-SJTU/AuxFormer.
****
Measuring Asymmetric Gradient Discrepancy in Parallel Continual Learning
Fan Lyu, Qing Sun, Fanhua Shang, Liang Wan, Wei Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11411-11420
In Parallel Continual Learning (PCL), the parallel multiple tasks start and end training unpredictably, thus suffering from training conflict and catastrophic forgetting issues. The two issues are raised because the gradients from parallel tasks differ in directions and magnitudes. Thus, in this paper, we formulate the PCL into a minimum distance optimization problem among gradients and propose an explicit Asymmetric Gradient Distance (AGD) to evaluate the gradient discrepancy in PCL. AGD considers both gradient magnitude ratios and directions, and has a tolerance when updating with a small gradient of inverse direction, which reduces the imbalanced influence of gradients on parallel task training. Moreover, we propose a novel Maximum Discrepancy Optimization (MaxDO) strategy to minimize the maximum discrepancy among multiple gradients. Solving by MaxDO with AGD, parallel training reduces the influence of the training conflict and suppresses the catastrophic forgetting of finished tasks. Extensive experiments validate the effectiveness of our approach on three image recognition datasets.
****
StyleLipSync: Style-based Personalized Lip-sync Video Generation
Taekyung Ki, Dongchan Min; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22841-22850

In this paper, we present StyleLipSync, a style-based personalized lip-sync video generative model that can generate identity-agnostic lip-synchronizing video from arbitrary audio. To generate a video of arbitrary identities, we leverage expressive lip prior from the semantically rich latent space of a pre-trained StyleGAN, where we can also design a video consistency with a linear transformation. In contrast to the previous lip-sync methods, we introduce pose-aware masking that dynamically locates the mask to improve the naturalness over frames by utilizing a 3D parametric mesh predictor frame by frame. Moreover, we propose a few-shot lip-sync adaptation method for an arbitrary person by introducing a sync regularizer that preserves lip-sync generalization while enhancing the person-specific visual information. Extensive experiments demonstrate that our model can generate accurate lip-sync videos even with the zero-shot setting and enhance characteristics of an unseen face using a few seconds of target video through the proposed adaptation method.
****
Cross Contrasting Feature Perturbation for Domain Generalization
Chenming Li, Daoan Zhang, Wenjian Huang, Jianguo Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1327-1337
Domain generalization (DG) aims to learn a robust model from source domains that generalize well on unseen target domains. Recent studies focus on generating novel domain samples or features to diversify distributions complementary to source domains. Yet, these approaches can hardly deal with the restriction that the samples synthesized from various domains can cause semantic distortion. In this paper, we propose a Cross Contrasting Feature Perturbation (CCFP) framework to simulate domain shift by generating perturbed features in the latent space while regularizing the model prediction against domain shift. Different from the previous fixed synthesizing strategy, we design modules with learnable feature perturbations and semantic consistency constraints. In contrast to prior work, our method does not use any generative-based models or domain labels. We conduct extensive experiments on a standard DomainBed benchmark with a strict evaluation protocol for a fair comparison. Comprehensive experiments show that our method outperforms the previous state-of-the-art, and quantitative analyses illustrate that our approach can alleviate the domain shift problem in out-of-distribution (OOD) scenarios.
****
DiffusionRet: Generative Text-Video Retrieval with Diffusion Model
Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, Jie Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2470-2481
Existing text-video retrieval solutions are, in essence, discriminant models focused on maximizing the conditional likelihood, i.e., p(candidates|query). While straightforward, this de facto paradigm overlooks the underlying data distribution p(query), which makes it challenging to identify out-of-distribution data. To address this limitation, we creatively tackle this task from a generative viewpoint and model the correlation between the text and the video as their joint probability p(candidates,query). This is accomplished through a diffusion-based text-video retrieval framework (DiffusionRet), which models the retrieval task as a process of gradually generating joint distribution from noise. During training, DiffusionRet is optimized from both the generation and discrimination perspectives, with the generator being optimized by generation loss and the feature extractor trained with contrastive loss. In this way, DiffusionRet cleverly leverages the strengths of both generative and discriminative methods. Extensive experiments on five commonly used text-video retrieval benchmarks, including MSRVTT, LSMDC, MSVD, ActivityNet Captions, and DiDeMo, with superior performances, justify the efficacy of our method. More encouragingly, without any modification, DiffusionRet even performs well in out-domain retrieval settings. We believe this work brings fundamental insights into the related fields. Code is available at https://github.com/jpthu17/DiffusionRet.
****
Efficient 3D Semantic Segmentation with Superpoint Transformer

Damien Robert, Hugo Raguet, Loic Landrieu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17195-17204
We introduce a novel superpoint-based transformer architecture for efficient semantic segmentation of large-scale 3D scenes. Our method incorporates a fast algorithm to partition point clouds into a hierarchical superpoint structure, which makes our preprocessing 7 times faster than existing superpoint-based approaches. Additionally, we leverage a self-attention mechanism to capture the relationships between superpoints at multiple scales, leading to state-of-the-art performance on three challenging benchmark datasets: S3DIS (76.0% mIoU 6-fold validation), KITTI-360 (63.5% on Val), and DALES (79.6%). With only 212k parameters, our approach is up to 200 times more compact than other state-of-the-art models while maintaining similar performance. Furthermore, our model can be trained on a single GPU in 3 hours for a fold of the S3DIS dataset, which is 7x to 70x fewer GPU-hours than the best-performing methods. Our code and models are accessible at github.com/drprojects/superpoint_transformer.
****
Adversarial Finetuning with Latent Representation Constraint to Mitigate Accuracy-Robustness Tradeoff
Satoshi Suzuki, Shin'ya Yamaguchi, Shoichiro Takeda, Sekitoshi Kanai, Naoki Makishima, Atsushi Ando, Ryo Masumura; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4390-4401
This paper addresses the tradeoff between standard accuracy on clean examples and robustness against adversarial examples in deep neural networks (DNNs).
Although adversarial training (AT) improves robustness, it degrades the standard accuracy, thus yielding the tradeoff.
To mitigate this tradeoff, we propose a novel AT method called ARREST, which comprises three components: (i) adversarial finetuning (AFT), (ii) representation-guided knowledge distillation (RGKD), and (iii) noisy replay (NR).
AFT trains a DNN on adversarial examples by initializing its parameters with a DNN that is standardly pretrained on clean examples.
RGKD and NR respectively entail a regularization term and an algorithm to preserve latent representations of clean examples during AFT.
RGKD penalizes the distance between the representations of the standardly pretrained and AFT DNNs.
NR switches input adversarial examples to nonadversarial ones when the representation changes significantly during AFT.
By combining these components, ARREST achieves both high standard accuracy and robustness.
Experimental results demonstrate that ARREST mitigates the tradeoff more effectively than previous AT-based methods do.
****
HyperDiffusion: Generating Implicit Neural Fields with Weight-Space Diffusion
Ziya Erkoç, Fangchang Ma, Qi Shan, Matthias Nießner, Angela Dai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14300-14310
Implicit neural fields, typically encoded by a multilayer perceptron (MLP) that maps from coordinates (e.g., xyz) to signals (e.g., signed distances), have shown remarkable promise as a high-fidelity and compact representation. However, the lack of a regular and explicit grid structure also makes it challenging to apply generative modeling directly on implicit neural fields in order to synthesize new data. To this end, we propose HyperDiffusion, a novel approach for unconditional generative modeling of implicit neural fields. HyperDiffusion operates directly on MLP weights and generates new neural implicit fields encoded by synthesized MLP parameters. Specifically, a collection of MLPs is first optimized to faithfully represent individual data samples. Subsequently, a diffusion process is trained in this MLP weight space to model the underlying distribution of neural implicit fields. HyperDiffusion enables diffusion modeling over a implicit, compact, and yet high-fidelity representation of complex signals across various dimensionalities within one single unified framework.
Experiments on both 3D shapes and 4D mesh animations demonstrate the effectiven

ess of our approach with significant improvement over prior work in high-fidelit
y synthesis.
****
Retinexformer: One-stage Retinex-based Transformer for Low-light Image Enhanceme
nt
Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, Yulun Zhang; Procee
dings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023,
pp. 12504-12513
When enhancing low-light images, many deep learning algorithms are based on the
Retinex theory. However, the Retinex model does not consider the corruptions hid
den in the dark or introduced by the light-up process. Besides, these methods us
ually require a tedious multi-stage training pipeline and rely on convolutional
neural networks, showing limitations in capturing long-range dependencies. In th
is paper, we formulate a simple yet principled One-stage Retinex-based Framework
 (ORF). ORF first estimates the illumination information to light up the low-lig
ht image and then restores the corruption to produce the enhanced image. We desi
gn an Illumination-Guided Transformer (IGT) that utilizes illumination represent
ations to direct the modeling of non-local interactions of regions with differen
t lighting conditions. By plugging IGT into ORF, we obtain our algorithm, Retine
xformer. Comprehensive quantitative and qualitative experiments demonstrate that
 our Retinexformer significantly outperforms state-of-the-art methods on thirtee
n benchmarks. The user study and application on low-light object detection also
reveal the latent practical values of our method. Code is available at https://g
ithub.com/caiyuanhao1998/Retinexformer
****
Minimum Latency Deep Online Video Stabilization
Zhuofan Zhang, Zhen Liu, Ping Tan, Bing Zeng, Shuaicheng Liu; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23030-23
039
We present a novel camera path optimization framework for the task of online vid
eo stabilization. Typically, a stabilization pipeline consists of three steps: m
otion estimating, path smoothing, and novel view rendering. Most previous method
s concentrate on motion estimation, proposing various global or local motion mod
els. In contrast, path optimization receives relatively less attention, especial
ly in the important online setting, where no future frames are available. In thi
s work, we adopt recent off-the-shelf high-quality deep motion models for motion
 estimation to recover the camera trajectory and focus on the latter two steps.
Our network takes a short 2D camera path in a sliding window as input and output
s the stabilizing warp field of the last frame in the window, which warps the co
ming frame to its stabilized position. A hybrid loss is well-defined to constrai
n the spatial and temporal consistency. In addition, we build a motion dataset t
hat contains stable and unstable motion pairs for the training. Extensive experi
ments demonstrate that our approach significantly outperforms state-of-the-art o
nline methods both qualitatively and quantitatively and achieves comparable perf
ormance to offline methods.
****
Speech2Lip: High-fidelity Speech to Lip Generation by Learning from a Short Vide
o
Xiuzhe Wu, Pengfei Hu, Yang Wu, Xiaoyang Lyu, Yan-Pei Cao, Ying Shan, Wenming Ya
ng, Zhongqian Sun, Xiaojuan Qi; Proceedings of the IEEE/CVF International Confer
ence on Computer Vision (ICCV), 2023, pp. 22168-22177
Synthesizing realistic videos according to a given speech is still an open chall
enge. Previous works have been plagued by issues such as inaccurate lip shape ge
neration and poor image quality. The key reason is that only motions and appeara
nces on limited facial areas (e.g., lip area) are mainly driven by the input spe
ech. Therefore, directly learning a mapping function from speech to the entire h
ead image is prone to ambiguity, particularly when using a short video for train
ing. We thus propose a decomposition-synthesis-composition framework named Speec
h to Lip (Speech2Lip) that disentangles speech-sensitive and speech-insensitive
motion/appearance to facilitate effective learning from limited training data, r

esulting in the generation of natural-looking videos. First, given a fixed head pose (i.e., canonical space), we present a speech-driven implicit model for lip image generation which concentrates on learning speech-sensitive motion and appearance. Next, to model the major speech-insensitive motion (i.e., head movement), we introduce a geometry-aware mutual explicit mapping (GAMEM) module that establishes geometric mappings between different head poses. This allows us to paste generated lip images at the canonical space onto head images with arbitrary poses and synthesize talking videos with natural head movements. In addition, a Blend-Net and a contrastive sync loss are introduced to enhance the overall synthesis performance. Quantitative and qualitative results on three benchmarks demonstrate that our model can be trained by a video of just a few minutes in length and achieve state-of-the-art performance in both visual quality and speech-visual synchronization. Code: https://github.com/CVMI-Lab/Speech2Lip.
****
UHDNeRF: Ultra-High-Definition Neural Radiance Fields
Quewei Li, Feichao Li, Jie Guo, Yanwen Guo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23097-23108
We propose UHDNeRF, a new framework for novel view synthesis on the challenging ultra-high-resolution (e.g., 4K) real-world scenes. Previous NeRF methods are not specifically designed for rendering on extremely high resolutions, leading to burry results with notable detail-losing problems even though trained on 4K images. This is mainly due to the mismatch between the high-resolution inputs and the low-dimensional volumetric representation. To address this issue, we introduce an adaptive implicit-explicit scene representation with which an explicit sparse point cloud is used to boost the performance of an implicit volume on modeling subtle details. Specifically, we reconstruct the complex real-world scene with a frequency separation strategy that the implicit volume learns to represent the low-frequency properties of the whole scene, and the sparse point cloud is used for reproducing high-frequency details. To better explore the information embedded in the point cloud, we extract a global structure feature and a local point-wise feature from the point cloud for each sample located in the high-frequency regions. Furthermore, a patch-based sampling strategy is introduced to reduce the computational cost. The high-fidelity rendering results demonstrate the superiority of our method for retaining high-frequency details at 4K ultra-high-resolution scenarios against state-of-the-art NeRF-based solutions.
****
Linear Spaces of Meanings: Compositional Structures in Vision-Language Models
Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, Stefano Soatto; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15395-15404
We investigate compositional structures in data embeddings from pre-trained vision-language models (VLMs). Traditionally, compositionality has been associated with algebraic operations on embeddings of words from a pre-existing vocabulary. In contrast, we seek to approximate representations from an encoder as combinations of a smaller set of vectors in the embedding space. These vectors can be seen as "ideal words" for generating concepts directly within embedding space of the model. We first present a framework for understanding compositional structures from a geometric perspective. We then explain what these compositional structures entail probabilistically in the case of VLM embeddings, providing intuitions for why they arise in practice. Finally, we empirically explore these structures in CLIP's embeddings and we evaluate their usefulness for solving different vision-language tasks such as classification, debiasing, and retrieval. Our results show that simple linear algebraic operations on embedding vectors can be used as compositional and interpretable methods for regulating the behavior of VLMs.
****
MULLER: Multilayer Laplacian Resizer for Vision
Zhengzhong Tu, Peyman Milanfar, Hossein Talebi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6877-6887
Image resizing operation is a fundamental preprocessing module in modern computer vision. Throughout the deep learning revolution, researchers have overlooked t

he potential of alternative resizing methods beyond the commonly used resizers that are readily available, such as nearest-neighbors, bilinear, and bicubic. The key question of our interest is whether the front-end resizer affects the performance of deep vision models? In this paper, we present an extremely lightweight multilayer Laplacian resizer with only a handful of trainable parameters, dubbed MULLER resizer. MULLER has a bandpass nature in that it learns to boost details in certain frequency subbands that benefit the downstream recognition models. We show that MULLER can be easily plugged into various training pipelines, and it effectively boosts the performance of the underlying vision task with little to no extra cost. Specifically, we select a state-of-the-art vision Transformer, MaxViT, as the baseline, and show that, if trained with MULLER, MaxViT gains up to 0.6% top-1 accuracy, and meanwhile enjoys 36% inference cost saving to achieve similar top-1 accuracy on ImageNet-1k, as compared to the standard training scheme. Notably, MULLER's performance also scales with model size and training data size such as ImageNet-21k and JFT, and it is widely applicable to multiple vision tasks, including image classification, object detection and segmentation, as well as image quality assessment.
 The code is available at https://github.com/google-research/google-research/tree/master/muller.
****
X-VoE: Measuring eXplanatory Violation of Expectation in Physical Events
Bo Dai, Linge Wang, Baoxiong Jia, Zeyu Zhang, Song-Chun Zhu, Chi Zhang, Yixin Zhu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3992-4002
Intuitive physics is pivotal for human understanding of the physical world, enabling prediction and interpretation of events even in infancy. Nonetheless, replicating this level of intuitive physics in artificial intelligence (AI) remains a formidable challenge. This study introduces X-VoE, a comprehensive benchmark dataset, to assess AI agents' grasp of intuitive physics. Built on the developmental psychology-rooted Violation of Expectation (VoE) paradigm, X-VoE establishes a higher bar for the explanatory capacities of intuitive physics models. Each VoE scenario within X-VoE encompasses three distinct settings, probing models' comprehension of events and their underlying explanations. Beyond model evaluation, we present an explanation-based learning system that captures physics dynamics and infers occluded object states solely from visual sequences, without explicit occlusion labels. Experimental outcomes highlight our model's alignment with human commonsense when tested against X-VoE. A remarkable feature is our model's ability to visually expound VoE events by reconstructing concealed scenes. Concluding, we discuss the findings' implications and outline future research directions. Through X-VoE, we catalyze the advancement of AI endowed with human-like intuitive physics capabilities.
****
Tracking by Natural Language Specification with Long Short-term Context Decoupling
Ding Ma, Xiangqian Wu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14012-14021
The main challenge of Tracking by Natural Language Specification (TNL) is to predict the movement of the target object by giving two heterogeneous information, e.g., one is the static description of the main characteristics of a video contained in the textual query, i.e., long-term context; the other one is an image patch containing the object and its surroundings cropped from the current frame, i.e., the search area. Currently, most methods still struggle with the rationality of using those two information and simply fusing the two. However, the linguistic information contained in the textual query and the visual representation stored in the search area may sometimes be inconsistent, in which case the direct fusion of the two may lead to conflicts. To address this problem, we propose DecoupleTNL, introducing a video clip containing short-term context information into the framework of TNL and exploring a proper way to reduce the impact when visual representation is inconsistent with linguistic information. Concretely, we design two jointly optimized tasks, i.e., short-term context-matching and long-term

context-perceiving. The context-matching task aims to gather the dynamic short-term context information in a period, while the context-perceiving task tends to extract the static long-term context information. After that, we design a long short-term modulation module to integrate both context information for accurate tracking. Extensive experiments have been conducted on three tracking benchmark datasets to demonstrate the superiority of DecoupleTNL
****
COOP: Decoupling and Coupling of Whole-Body Grasping Pose Generation
Yanzhao Zheng, Yunzhou Shi, Yuhao Cui, Zhongzhou Zhao, Zhiling Luo, Wei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2163-2173
Generating life-like whole-body human grasping has garnered significant attention in the field of computer graphics. Existing works have demonstrated the effectiveness of keyframe-guided motion generation framework, witch focus on modeling the grasping motions of humans in temporal sequence when the target objects are placed in front of them. However, the generated grasping poses of the human body in the key-frames are limited, failing to capture the full range of grasping poses that humans are capable of.
To address this issue, we propose a novel framework called COOP (DeCOupling and COupling of Whole-Body GrasPing Pose Generation) to synthesize life-like whole-body poses that cover the widest range of human grasping capabilities. In this framework, we first decouple the whole-body pose into body pose and hand pose and model them separately, which allows us to pre-train the body model with out-of-domain data easily. Then, we couple these two generated body parts through a unified optimization algorithm.
Furthermore, we design a simple evaluation method to evaluate the generalization ability of models in generating grasping poses for objects placed at different positions. The experimental results demonstrate the efficacy and superiority of our method. And COOP holds great potential as a plug-and-play component for other domains in whole-body pose generation. Our models and code are available at https://github.com/zhengyanzhao1997/COOP.
****
Pyramid Dual Domain Injection Network for Pan-sharpening
Xuanhua He, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, Man Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12908-12917
Pan-sharpening, a panchromatic image guided low-spatial-resolution multi-spectral super-resolution task, aims to reconstruct the missing high-frequency information of high-resolution multi-spectral counterpart. Although the inborn connection with frequency domain, existing pan-sharpening research has almost investigated the potential solution upon frequency domain, thus limiting the model performance improvement. To this end, we first revisit the degradation process of pan-sharpening in Fourier space, and then devise a Pyramid Dual Domain Injection Pan-sharpening Network upon the above observation by fully exploring and exploiting the distinguished information in both the spatial and frequency domains. Specifically, the proposed network is organized with multi-scale U-shape manner and composed by two core parts: a spatial guidance pyramid sub-network for fusing local spatial information and a frequency guidance pyramid sub-network for fusing global frequency domain information, thus encouraging dual-domain complementary learning. In this way, the model can capture multi-scale dual-domain information to enable generating high-quality pan-sharpening results. Quantitative and qualitative experiments over multiple datasets demonstrate that our method performs the best against other state-of-the-art ones and comprises a strong generalization ability for real-world scenes.
****
Why do networks have inhibitory/negative connections?
Qingyang Wang, Mike A. Powell, Ali Geisa, Eric Bridgeford, Carey E. Priebe, Joshua T. Vogelstein; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22551-22559
Why do brains have inhibitory connections? Why do deep networks have negative we

ights? We propose an answer from the perspective of representation capacity. We believe representing functions is the primary role of both (i) the brain in natural intelligence, and (ii) deep networks in artificial intelligence. Our answer to why there are inhibitory/negative weights is: to learn more functions. We prove that, in the absence of negative weights, neural networks with non-decreasing activation functions are not universal approximators. While this may be an intuitive result to some, to the best of our knowledge, there is no formal theory, in either machine learning or neuroscience, that demonstrates why negative weights are crucial in the context of representation capacity. Further, we provide insights on the geometric properties of the representation space that non-negative deep networks cannot represent. We expect these insights will yield a deeper understanding of more sophisticated inductive priors imposed on the distribution of weights that lead to more efficient biological and machine learning.
****
Ordinal Label Distribution Learning
Changsong Wen, Xin Zhang, Xingxu Yao, Jufeng Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23481-23491
Label distribution learning (LDL) is a recent hot topic, in which ambiguity is modeled via description degrees of the labels. However, in common LDL tasks, e.g., age estimation, labels are in an intrinsic order. The conventional LDL paradigm adopts a per-label manner for optimization, neglecting the internal sequential patterns of labels. Therefore, we propose a new paradigm, termed ordinal label distribution learning (OLDL). We model the sequential patterns of labels from aspects of spatial, semantic, and temporal order relationships. The spatial order depicts the relative position between arbitrary labels. We build cross-label transformation between distributions, which is determined by the spatial margin in labels. Labels naturally yield different semantics, so the semantic order is represented by constructing semantic correlations between arbitrary labels. The temporal order describes that the presence of labels is determined by their order, i.e. five after four. The value of a particular label contains information about previous labels, and we adopt cumulative distribution to construct this relationship. Based on these characteristics of ordinal labels, we propose the learning objectives and evaluation metrics for OLDL, namely CAD, QFD, and CJS. Comprehensive experiments conducted on four tasks demonstrate the superiority of OLDL against other existing LDL methods in both traditional and newly proposed metrics. Our project page can be found at https://downdric23.github.io/.
****
Model Calibration in Dense Classification with Adaptive Label Perturbation
Jiawei Liu, Changkun Ye, Shan Wang, Ruikai Cui, Jing Zhang, Kaihao Zhang, Nick Barnes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1173-1184
For safety-related applications, it is crucial to produce trustworthy deep neural networks whose prediction is associated with confidence that can represent the likelihood of correctness for subsequent decision-making. Existing dense binary classification models are prone to being over-confident. To improve model calibration, we propose Adaptive Stochastic Label Perturbation (ASLP) which learns a unique label perturbation level for each training image. ASLP employs our proposed Self-Calibrating Binary Cross Entropy (SC-BCE) loss, which unifies label perturbation processes including stochastic approaches (like DisturbLabel), and label smoothing, to correct calibration while maintaining classification rates. ASLP follows Maximum Entropy Inference of classic statistical mechanics to maximise prediction entropy with respect to missing information. It performs this while: (1) preserving classification accuracy on known data as a conservative solution, or (2) specifically improves model calibration degree by minimising the gap between the prediction accuracy and expected confidence of target training label. Extensive results demonstrate that ASLP can significantly improve calibration degrees of dense binary classification models on both in-distribution and out-of-distribution data.
****
Boosting Multi-modal Model Performance with Adaptive Gradient Modulation

Hong Li, Xingyu Li, Pengbo Hu, Yinuo Lei, Chunxiao Li, Yi Zhou; Proceedings of t
he IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22214-
22224
While the field of multi-modal learning keeps growing fast, the deficiency of th
e standard joint training paradigm has become clear through recent studies. They
 attribute the sub-optimal performance of the jointly trained model to the modal
ity competition phenomenon. Existing works attempt to improve the jointly traine
d model by modulating the training process. Despite their effectiveness, those m
ethods can only apply to late fusion models. More importantly, the mechanism of
the modality competition remains unexplored. In this paper, we first propose an
adaptive gradient modulation method that can boost the performance of multi-moda
l models with various fusion strategies. Extensive experiments show that our met
hod surpasses all existing modulation methods. Furthermore, to have a quantitati
ve understanding of the modality competition and the mechanism behind the effect
iveness of our modulation method, we introduce a novel metric to measure the com
petition strength. This metric is built on the mono-modal concept, a function th
at is designed to represent the competition-less state of a modality. Through sy
stematic investigation, our results confirm the intuition that the modulation en
courages the model to rely on the more informative modality. In addition, we fin
d that the jointly trained model typically has a preferred modality on which the
 competition is weaker than other modalities. However, this preferred modality n
eed not dominate others. Our code will be available at https://github.com/lihong
2303/AGM_ICCV2023.
****
Semantic Information in Contrastive Learning
Shengjiang Quan, Masahiro Hirano, Yuji Yamakawa; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 5686-5696
This work investigates the functionality of Semantic information in Contrastive
Learning (SemCL). An advanced pretext task is designed: a contrast is performed
between each object and its environment, taken from a scene. This allows the Sem
CL pretrained model to extract objects from their environment in an image, signi
ficantly improving the spatial understanding of the pretrained models. Downstrea
m tasks of semantic/instance segmentation, object detection and depth estimation
 are implemented on PASCAl VOC, Cityscapes, COCO, KITTI, etc. SemCL pretrained m
odels substantially outperform ImageNet pretrained counterparts and are competit
ive with well-known works on downstream tasks. The results suggest that a dedica
ted pretext task leveraging semantic information can be powerful in benchmarks r
elated to spatial understanding. The code is available at https://github.com/sji
ang95/semcl.
****
Structure and Content-Guided Video Synthesis with Diffusion Models
Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, Anastasi
s Germanidis; Proceedings of the IEEE/CVF International Conference on Computer V
ision (ICCV), 2023, pp. 7346-7356
Text-guided generative diffusion models unlock powerful image creation and editi
ng tools. Recent approaches that edit the content of footage while retaining str
ucture require expensive re-training for every input or rely on error-prone prop
agation of image edits across frames. In this work, we present a structure and c
ontent-guided video diffusion model that edits videos based on descriptions of t
he desired output. Conflicts between user-provided content edits and structure r
epresentations occur due to insufficient disentanglement between the two aspects
. As a solution, we show that training on monocular depth estimates with varying
 levels of detail provides control over structure and content fidelity. A novel
guidance method, enabled by joint video and image training, exposes explicit con
trol over temporal consistency. Our experiments demonstrate a wide variety of su
ccesses; fine-grained control over output characteristics, customization based o
n a few reference images, and a strong user preference towards results by our mo
del.
****
NeSS-ST: Detecting Good and Stable Keypoints with a Neural Stability Score and t

he Shi-Tomasi detector
Konstantin Pakulev, Alexander Vakhitov, Gonzalo Ferrer; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9578-9588
Learning a feature point detector presents a challenge both due to the ambiguity of the definition of a keypoint and, correspondingly, the need for specially prepared ground truth labels for such points. In our work, we address both of these issues by utilizing a combination of a hand-crafted Shi-Tomasi detector, a specially designed metric that assesses the quality of keypoints, the stability score (SS), and a neural network. We build on the principled and localized keypoints provided by the Shi-Tomasi detector and learn the neural network to select good feature points via the stability score. The neural network incorporates the knowledge from the training targets in the form of the neural stability score (NeSS). Therefore, our method is named NeSS-ST since it combines the Shi-Tomasi detector and the properties of the neural stability score. It only requires sets of images for training without dataset pre-labeling or the need for reconstructed correspondence labels. We evaluate NeSS-ST on HPatches, ScanNet, MegaDepth and IMC-PT demonstrating state-of-the-art performance and good generalization on downstream tasks. The project repository is available at: https://github.com/KonstantinPakulev/NeSS-ST.
****
Beyond Skin Tone: A Multidimensional Measure of Apparent Skin Color
William Thong, Przemyslaw Joniak, Alice Xiang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4903-4913
This paper strives to measure apparent skin color in computer vision, beyond a unidimensional scale on skin tone. In their seminal paper Gender Shades, Buolamwini and Gebru have shown how gender classification systems can be biased against women with darker skin tones. Subsequently, fairness researchers and practitioners have adopted the Fitzpatick skin type classification as a common measure to assess skin color bias in computer vision systems. While effective, the Fitzpatick scale only focuses on the skin tone ranging from light to dark. Towards a more comprehensive measure of skin color, we introduce the hue angle ranging from red to yellow. When applied to images, the hue dimension reveals additional biases related to skin color in both computer vision datasets and models. We then recommend multidimensional skin color scales, relying on both skin tone and hue, for fairness assessments.
****
PODA: Prompt-driven Zero-shot Domain Adaptation
Mohammad Fahes, Tuan-Hung Vu, Andrei Bursuc, Patrick Pérez, Raoul de Charette; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18623-18633
Domain adaptation has been vastly investigated in computer vision but still requires access to target images at train time, which might be intractable in some uncommon conditions. In this paper, we propose the task of 'Prompt-driven Zero-shot Domain Adaptation', where we adapt a model trained on a source domain using only a general description in natural language of the target domain, i.e., a prompt. First, we leverage a pretrained contrastive vision-language model (CLIP) to optimize affine transformations of source features, steering them towards the target text embedding while preserving their content and semantics. To achieve this, we propose Prompt-driven Instance Normalization (PIN). Second, we show that these prompt-driven augmentations can be used to perform zero-shot domain adaptation for semantic segmentation. Experiments demonstrate that our method significantly outperforms CLIP-based style transfer baselines on several datasets for the downstream task at hand, even surpassing one-shot unsupervised domain adaptation. A similar boost is observed on object detection and image classification. The code is available at https://github.com/astra-vision/PODA .
****
Video Action Segmentation via Contextually Refined Temporal Keypoints
Borui Jiang, Yang Jin, Zhentao Tan, Yadong Mu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13836-13845
Video action segmentation refers to the task of densely casting each video frame

or short segment in an untrimmed video into some pre-specified action categorie
s. Although recent years have witnessed a great promise in the development of ac
tion segmentation techniques.A large body of existing methods still rely on fram
e-wise segmentation, which tends to render fragmentary results (i.e., over-segme
ntation).To effectively address above issues, we here propose a video action seg
mentation model that implements the novel idea of Refined Temporal Keypoints (RT
K) for overcoming caveats of existing methods.To act effectively, the proposed m
odel initially seeks for high-quality, sparse temporal keypoints by extracting n
on-local cues from the video, rather than conducting frame-wise classification a
s in many competing methods.Afterwards, large improvements over the inital tempo
ral keypoints are pin-pointed as contributions by further refining and re-assemb
ling operations. In specific, we develop a graph matching module that aggregates
structural information between different temporal keypoints by learning the cor
responding relationship of the temporal source graphs and the annotated target g
raphs. The initial temporal keypoints are refined by the encoded structural info
rmation reusing the graph matching module.A few set of prior rules are harnessed
for post-processing and re-assembling all temporal keypoints.The remaining temp
oral keypoiting going through all refinement are used to generate the final acti
on segmentation results.We perform experiments on three popular datasets: 50sala
ds, GTEA and Breakfast, and our methods significantly outperforms the current me
thods, particularly achieves the state-of-the-art F1@50 scores of 83.4%, 79.5%,
and 60.5% on three datasets, respectively.
****
Shatter and Gather: Learning Referring Image Segmentation with Text Supervision
Dongwon Kim, Namyup Kim, Cuiling Lan, Suha Kwak; Proceedings of the IEEE/CVF Int
ernational Conference on Computer Vision (ICCV), 2023, pp. 15547-15557
Referring image segmentation, the task of segmenting any arbitrary entities desc
ribed in free-form texts, opens up a variety of vision applications.
However, manual labeling of training data for this task is prohibitively costly
, leading to lack of labeled data for training. We address this issue by a weakl
y supervised learning approach using text descriptions of training images as the
only source of supervision. To this end, we first present a new model that disc
overs semantic entities in input image and then combines such entities relevant
to text query to predict the mask of the referent.
We also present a new loss function that allows the model to be trained without
any further supervision. Our method was evaluated on four public benchmarks for
referring image segmentation, where it clearly outperformed the existing method
for the same task and recent open-vocabulary segmentation models on all the ben
chmarks.
****
Two-in-One Depth: Bridging the Gap Between Monocular and Binocular Self-Supervis
ed Depth Estimation
Zhengming Zhou, Qiulei Dong; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2023, pp. 9411-9421
Monocular and binocular self-supervised depth estimations are two important and
related tasks in computer vision, which aim to predict scene depths from single
images and stereo image pairs respectively. In literature, the two tasks are usu
ally tackled separately by two different kinds of models, and binocular models g
enerally fail to predict depth from single images, while the prediction accuracy
of monocular models is generally inferior to binocular models. In this paper, w
e propose a Two-in-One self-supervised depth estimation network, called TiO-Dept
h, which could not only compatibly handle the two tasks, but also improve the pr
ediction accuracy. TiO-Depth employs a Siamese architecture and each sub-network
of it could be used as a monocular depth estimation model. For binocular depth
estimation, a Monocular Feature Matching module is proposed for incorporating th
e stereo knowledge between the two images, and the full TiO-Depth is used to pre
dict depths. We also design a multi-stage joint-training strategy for improving
the performances of TiO-Depth in both two tasks by combining the relative advant
ages of them. Experimental results on the KITTI, Cityscapes, and DDAD datasets d
emonstrate that TiO-Depth outperforms both the monocular and binocular state-of-

the-art methods in most cases, and further verify the feasibility of a two-in-on
e network for monocular and binocular depth estimation. The code is available at
 https://github.com/ZM-Zhou/TiO-Depth_pytorch.
****
SAFL-Net: Semantic-Agnostic Feature Learning Network with Auxiliary Plugins for
Image Manipulation Detection
Zhihao Sun, Haoran Jiang, Danding Wang, Xirong Li, Juan Cao; Proceedings of the
IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22424-224
33
Since image editing methods in real world scenarios cannot be exhausted, general
ization is a core challenge for image manipulation detection, which could be sev
erely weakened by semantically related features. In this paper we propose SAFL-N
et, which constrains a feature extractor to learn semantic-agnostic features by
designing specific modules with corresponding auxiliary tasks. Applying constrai
nts directly to the features extracted by the encoder helps it learn semantic-ag
nostic manipulation trace features, which prevents the biases related to semanti
c information within the limited training data and improves generalization capab
ilities. The consistency of auxiliary boundary prediction task and original regi
on prediction task is guaranteed by a feature transformation structure. Experime
nts on various public datasets and comparisons in multiple dimensions demonstrat
e that SAFL-Net is effective for image manipulation detection.
****
DataDAM: Efficient Dataset Distillation with Attention Matching
Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z. Liu, Yuri A. Lawryshyn, Konst
antinos N. Plataniotis; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2023, pp. 17097-17107
Researchers have long tried to minimize training costs in deep learning while ma
intaining strong generalization across diverse datasets. Emerging research on da
taset distillation aims to reduce training costs by creating a small synthetic s
et that contains the information of a larger real dataset and ultimately achieve
s test accuracy equivalent to a model trained on the whole dataset. Unfortunatel
y, the synthetic data generated by previous methods are not guaranteed to distri
bute and discriminate as well as the original training data, and they incur sign
ificant computational costs. Despite promising results, there still exists a sig
nificant performance gap between models trained on condensed synthetic sets and
those trained on the whole dataset. In this paper, we address these challenges u
sing efficient Dataset Distillation with Attention Matching (DataDAM), achieving
 state-of-the-art performance while reducing training costs. Specifically, we le
arn synthetic images by matching the spatial attention maps of real and syntheti
c data generated by different layers within a family of randomly initialized neu
ral networks. Our method outperforms the prior methods on several datasets, incl
uding CIFAR10/100, TinyImageNet, ImageNet-1K, and subsets of ImageNet-1K across
most of the settings, and achieves improvements of up to 6.5% and 4.1% on CIFAR1
00 and ImageNet-1K, respectively. We also show that our high-quality distilled i
mages have practical benefits for downstream applications, such as continual lea
rning and neural architecture search.
****
Rethinking Pose Estimation in Crowds: Overcoming the Detection Information Bottl
eneck and Ambiguity
Mu Zhou, Lucas Stoffl, Mackenzie Weygandt Mathis, Alexander Mathis; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14
689-14699
Frequent interactions between individuals are a fundamental challenge for pose e
stimation algorithms. Current pipelines either use an object detector together w
ith a pose estimator (top-down approach), or localize all body parts first and t
hen link them to predict the pose of individuals (bottom-up). Yet, when individu
als closely interact, top-down methods are ill-defined due to overlapping indivi
duals, and bottom-up methods often falsely infer connections to distant bodypart
s. Thus, we propose a novel pipeline called bottom-up conditioned top-down pose
estimation (BUCTD) that combines the strengths of bottom-up and top-down methods

. Specifically, we propose to use a bottom-up model as the detector, which in addition to an estimated bounding box provides a pose proposal that is fed as condition to an attention-based top-down model. We demonstrate the performance and efficiency of our approach on animal and human pose estimation benchmarks. On CrowdPose and OCHuman, we outperform previous state-of-the-art models by a significant margin. We achieve 78.5 AP on CrowdPose and 48.5 AP on OCHuman, an improvement of 8.6% and 7.8% over the prior art, respectively. Furthermore, we show that our method strongly improves the performance on multi-animal benchmarks involving fish and monkeys. The code is available at https://github.com/amathislab/BUCTD.
****
Social Diffusion: Long-term Multiple Human Motion Anticipation
Julian Tanke, Linguang Zhang, Amy Zhao, Chengcheng Tang, Yujun Cai, Lezi Wang, Po-Chen Wu, Juergen Gall, Cem Keskin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9601-9611
We propose Social Diffusion, a novel method for short-term and long-term forecasting of the motion of multiple persons as well as their social interactions.
Jointly forecasting motions for multiple persons involved in social activities is inherently a challenging problem due to the interdependencies between individuals.
In this work, we leverage a diffusion model conditioned on motion histories and causal temporal convolutional networks to forecast individually and contextually plausible motions for all participants. The contextual plausibility is achieved via an order-invariant aggregation function. As a second contribution, we design a new evaluation protocol that measures the plausibility of social interactions which we evaluate on the Haggling dataset, which features a challenging social activity where people are actively taking turns to talk and switching their attention.
We evaluate our approach on four datasets for multi-person forecasting where our approach outperforms the state-of-the-art in terms of motion realism and contextual plausibility.
****
Synchronize Feature Extracting and Matching: A Single Branch Framework for 3D Object Tracking
Teli Ma, Mengmeng Wang, Jimin Xiao, Huifeng Wu, Yong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9953-9963
Siamese network has been a de facto benchmark framework for 3D LiDAR object tracking with a shared-parametric encoder extracting features from template and search region, respectively. This paradigm relies heavily on an additional matching network to model the cross-correlation/similarity of the template and search region. In this paper, we forsake the conventional Siamese paradigm and propose a novel single-branch framework, SyncTrack, synchronizing the feature extracting and matching to avoid forwarding encoder twice for template and search region as well as introducing extra parameters of matching network. The synchronization mechanism is based on the dynamic affinity of the Transformer, and an in-depth analysis of the relevance is provided theoretically. Moreover, based on the synchronization, we introduce a novel Attentive Points-Sampling strategy into the Transformer layers (APST), replacing the random/Farthest Points Sampling (FPS) method with sampling under the supervision of attentive relations between the template and search region. It implies connecting point-wise sampling with the feature learning, beneficial to aggregating more distinctive and geometric features for tracking with sparse points. Extensive experiments on two benchmark datasets (KITTI and NuScenes) show that SyncTrack achieves state-of-the-art performance in real-time tracking.
****
Leveraging Intrinsic Properties for Non-Rigid Garment Alignment
Siyou Lin, Boyao Zhou, Zerong Zheng, Hongwen Zhang, Yebin Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14485-14496
We address the problem of aligning real-world 3D data of garments, which benefit

s many applications such as texture learning, physical parameter estimation, generative modeling of garments, etc. Existing extrinsic methods typically perform non-rigid iterative closest point and struggle to align details due to incorrect closest matches and rigidity constraints. While intrinsic methods based on functional maps can produce high-quality correspondences, they work under isometric assumptions and become unreliable for garment deformations which are highly non-isometric. To achieve wrinkle-level as well as texture-level alignment, we present a novel coarse-to-fine two-stage method that leverages intrinsic manifold properties with two neural deformation fields, in the 3D space and the intrinsic space, respectively. The coarse stage performs a 3D fitting, where we leverage intrinsic manifold properties to define a manifold deformation field. The coarse fitting then induces a functional map that produces an alignment of intrinsic embeddings. We further refine the intrinsic alignment with a second neural deformation field for higher accuracy. We evaluate our method with our captured garment dataset, GarmCap. The method achieves accurate wrinkle-level and texture-level alignment and works for difficult garment types such as long coats. Our project page is https://jsnln.github.io/iccv2023_intrinsic/index.html.
****
NeILF++: Inter-Reflectable Light Fields for Geometry and Material Estimation
Jingyang Zhang, Yao Yao, Shiwei Li, Jingbo Liu, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3601-3610
We present a novel differentiable rendering framework for joint geometry, material, and lighting estimation from multi-view images. In contrast to previous methods which assume a simplified environment map or co-located flashlights, in this work, we formulate the lighting of a static scene as one neural incident light field (NeILF) and one outgoing neural radiance field (NeRF). The key insight of the proposed method is the union of the incident and outgoing light fields through physically-based rendering and inter-reflections between surfaces, making it possible to disentangle the scene geometry, material, and lighting from image observations in a physically-based manner. The proposed incident light and inter-reflection framework can be easily applied to other NeRF systems. We show that our method can not only decompose the outgoing radiance into incident lights and surface materials, but also serve as a surface refinement module that further improves the reconstruction detail of the neural surface. We demonstrate on several datasets that the proposed method is able to achieve state-of-the-art results in terms of the geometry reconstruction quality, material estimation accuracy, and the fidelity of novel view rendering.
****
MAGI: Multi-Annotated Explanation-Guided Learning
Yifei Zhang, Siyi Gu, Yuyang Gao, Bo Pan, Xiaofeng Yang, Liang Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1977-1987
Explanation supervision is a technique in which the model is guided by human-generated explanations during training. This technique aims to improve the predictability of the model by incorporating human understanding of the prediction process into the training phase. This is a challenging task since it relies on the accuracy of human annotation labels. To obtain high-quality explanation annotations, using multiple annotations to do explanation supervision is a reasonable method. However, how to use multiple annotations to improve accuracy is particularly challenging due to the following: 1) The noisiness of annotations from different annotators; 2) The lack of pre-given information about the corresponding relationship between annotations and annotators; 3) Missing annotations since some images are not labeled by all annotators. To solve these challenges, we propose a Multi-annotated explanation-guided learning (MAGI) framework to do explanation supervision with comprehensive and high-quality generated annotations. We first propose a novel generative model to generate annotations from all annotators and infer them using a newly proposed variational inference-based technique by learning the characteristics of each annotator. We also incorporate an alignment mechanism into the generative model to infer the correspondence between annotations

and annotators in the training process. Extensive experiments on two datasets fr
om the medical imaging domain demonstrate the effectiveness of our proposed fram
ework in handling noisy annotations while obtaining superior prediction performa
nce compared with previous SOTA.
****
Adaptive Positional Encoding for Bundle-Adjusting Neural Radiance Fields
Zelin Gao, Weichen Dai, Yu Zhang; Proceedings of the IEEE/CVF International Conf
erence on Computer Vision (ICCV), 2023, pp. 3284-3294
Neural Radiance Fields have shown great potential to synthesize novel views with
 only a few discrete image observations of the world. However, the requirement o
f accurate camera parameters to learn scene representations limits its further a
pplication. In this paper, we present adaptive positional encoding (APE) for bun
dle-adjusting neural radiance fields to reconstruct the neural radiance fields f
rom unknown camera poses (or even intrinsics). Inspired by Fourier series regres
sion, we investigate its relationship with the positional encoding method and th
erefore propose APE where all frequency bands are trainable. Furthermore, we int
roduce period-activated multilayer perceptrons (PMLPs) to construct the implicit
 network for the high-order scene representations and fine-grain gradients durin
g backpropagation. Experimental results on public datasets demonstrate that the
proposed method with APE and PMLPs can outperform the state-of-the-art methods i
n accurate camera poses and high-fidelity view synthesis.
****
Inducing Neural Collapse to a Fixed Hierarchy-Aware Frame for Reducing Mistake S
everity
Tong Liang, Jim Davis; Proceedings of the IEEE/CVF International Conference on C
omputer Vision (ICCV), 2023, pp. 1443-1452
There is a recently discovered and intriguing phenomenon called Neural Collapse:
 at the terminal phase of training a deep neural network for classification, the
 within-class penultimate feature means and the associated classifier vectors of
 all flat classes collapse to the vertices of a simplex Equiangular Tight Frame
(ETF). Recent work has tried to exploit this phenomenon by fixing the related cl
assifier weights to a pre-computed ETF to induce neural collapse and maximize th
e separation of the learned features when training with imbalanced data. In this
 work, we propose to fix the linear classifier of a deep neural network to a Hie
rarchy-Aware Frame (HAFrame), instead of an ETF, and use a cosine similarity-bas
ed auxiliary loss to learn hierarchy-aware penultimate features that collapse to
 the HAFrame. We demonstrate that our approach reduces the mistake severity of t
he model's predictions while maintaining its top-1 accuracy on several datasets
of varying scales with hierarchies of heights ranging from 3 to 12. Code: https:
//github.com/ltong1130ztr/HAFrame.
****
PlanarTrack: A Large-scale Challenging Benchmark for Planar Object Tracking
Xinran Liu, Xiaoqiong Liu, Ziruo Yi, Xin Zhou, Thanh Le, Libo Zhang, Yan Huang,
Qing Yang, Heng Fan; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2023, pp. 20449-20458
Planar object tracking is a critical computer vision problem and has drawn incre
asing interest owing to its key roles in robotics, augmented reality, etc. Despi
te rapid progress, its further development, especially in the deep learning era,
 is largely hindered due to the lack of large-scale challenging benchmarks. Addr
essing this, we introduce PlanarTrack, a large-scale challenging planar tracking
 benchmark. Specifically, PlanarTrack consists of 1,000 videos with more than 49
0K images. All these videos are collected in complex unconstrained scenarios fro
m the wild, which makes PlanarTrack, compared with existing benchmarks, more cha
llenging but realistic for real-world applications. To ensure the high-quality a
nnotation, each frame in PlanarTrack is manually labeled using four corners with
 multiple-round careful inspection and refinement. To our best knowledge, Planar
Track, to date, is the largest and most challenging dataset dedicated to planar
object tracking. In order to analyze the proposed PlanarTrack, we evaluate 10 pl
anar trackers and conduct comprehensive comparisons and in-depth analysis. Our r
esults, not surprisingly, demonstrate that current top-performing planar tracker

s degenerate significantly on the challenging PlanarTrack and more efforts are needed to improve planar tracking in the future. In addition, we further derive a variant named PlanarTrack_BB for generic object tracking from PlanarTrack. Our evaluation of 10 excellent generic trackers on PlanarTrack_BB manifests that, surprisingly, PlanarTrack_BB is even more challenging than several popular generic tracking benchmarks and more attention should be paid to handle such planar objects, though they are rigid. All benchmarks and evaluations will be released.
****
Factorized Inverse Path Tracing for Efficient and Accurate Material-Lighting Estimation
Liwen Wu, Rui Zhu, Mustafa B. Yaldiz, Yinhao Zhu, Hong Cai, Janarbek Matai, Fatih Porikli, Tzu-Mao Li, Manmohan Chandraker, Ravi Ramamoorthi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3848-3858
Inverse path tracing has recently been applied to joint material and lighting estimation, given geometry and multi-view HDR observations of an indoor scene. However, it has two major limitations: path tracing is expensive to compute, and ambiguities exist between reflection and emission. Our Factorized Inverse Path Tracing (FIPT) addresses these challenges by using a factored light transport formulation and finds emitters driven by rendering errors. Our algorithm enables accurate material and lighting optimization faster than previous work, and is more effective at resolving ambiguities. The exhaustive experiments on synthetic scenes show that our method (1) outperforms state-of-the-art indoor inverse rendering and relighting methods particularly in the presence of complex illumination effects; (2) speeds up inverse path tracing optimization to less than an hour. We further demonstrate robustness to noisy inputs through material and lighting estimates that allow plausible relighting in a real scene. The source code is available at: https://github.com/lwwu2/fipt
****
P2C: Self-Supervised Point Cloud Completion from Single Partial Clouds
Ruikai Cui, Shi Qiu, Saeed Anwar, Jiawei Liu, Chaoyue Xing, Jing Zhang, Nick Barnes; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14351-14360
Point cloud completion aims to recover the complete shape based on a partial observation. Existing methods require either complete point clouds or multiple partial observations of the same object for learning. In contrast to previous approaches, we present Partial2Complete (P2C), the first self-supervised framework that completes point cloud objects using training samples consisting of only a single incomplete point cloud per object. Specifically, our framework groups incomplete point clouds into local patches as input and predicts masked patches by learning prior information from different partial objects. We also propose Region-Aware Chamfer Distance to regularize shape mismatch without limiting completion capability, and devise the Normal Consistency Constraint to incorporate a local planarity assumption, encouraging the recovered shape surface to be continuous and complete. In this way, P2C no longer needs multiple observations or complete point clouds as ground truth. Instead, structural cues are learned from a category-specific dataset to complete partial point clouds of objects. We demonstrate the effectiveness of our approach on both synthetic ShapeNet data and real-world ScanNet data, showing that P2C produces comparable results to methods trained with complete shapes, and outperforms methods learned with multiple partial observations. Code is available at https://github.com/CuiRuikai/Partial2Complete.
****
Overwriting Pretrained Bias with Finetuning Data
Angelina Wang, Olga Russakovsky; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3957-3968
Transfer learning is beneficial by allowing the expressive features of models pretrained on large-scale datasets to be finetuned for the target task of smaller, more domain-specific datasets. However, there is a concern that these pretrained models may come with their own biases which would propagate into the finetuned model. In this work, we investigate bias when conceptualized as both spurious c

orrelations between the target task and a sensitive attribute as well as underre presentation of a particular group in the dataset. Under both notions of bias, w e find that (1) models finetuned on top of pretrained models can indeed inherit their biases, but (2) this bias can be corrected for through relatively minor in terventions to the finetuning dataset, and often with a negligible impact to per formance. Our findings imply that careful curation of the finetuning dataset is important for reducing biases on a downstream task, and doing so can even compen sate for bias in the pretrained model.
****
Anti-DreamBooth: Protecting Users from Personalized Text-to-image Synthesis
Thanh Van Le, Hao Phung, Thuan Hoang Nguyen, Quan Dao, Ngoc N. Tran, Anh Tran; P roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2 023, pp. 2116-2127
Text-to-image diffusion models are nothing but a revolution, allowing anyone, ev en without design skills, to create realistic images from simple text inputs. Wi th powerful personalization tools like DreamBooth, they can generate images of a specific person just by learning from his/her few reference images. However, wh en misused, such a powerful and convenient tool can produce fake news or disturb ing content targeting any individual victim, posing a severe negative social imp act. In this paper, we explore a defense system called Anti-DreamBooth against s uch malicious use of DreamBooth. The system aims to add subtle noise perturbatio n to each user's image before publishing in order to disrupt the generation qual ity of any DreamBooth model trained on these perturbed images. We investigate a wide range of algorithms for perturbation optimization and extensively evaluate them on two facial datasets over various text-to-image model versions. Despite t he complicated formulation of DreamBooth and Diffusion-based text-to-image model s, our methods effectively defend users from the malicious use of those models. Their effectiveness withstands even adverse conditions, such as model or prompt/ term mismatching between training and testing. Our code will be available at htt ps://github.com/VinAIResearch/Anti-DreamBooth
****
Contrastive Continuity on Augmentation Stability Rehearsal for Continual Self-Su pervised Learning
Haoyang Cheng, Haitao Wen, Xiaoliang Zhang, Heqian Qiu, Lanxiao Wang, Hongliang Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2023, pp. 5707-5717
Self-supervised learning has attracted a lot of attention recently, which is abl e to learn powerful representations without any manual annotations. However, sel f-supervised learning needs to develop the ability to continuously learn to cope with a variety of real-world challenges, i.e., Continual Self-Supervised Learni ng (CSSL). Catastrophic forgetting is a notorious problem in CSSL, where the mod el tends to forget the learned knowledge. In practice, simple rehearsal or regul arization will bring extra negative effects while alleviating catastrophic forge tting in CSSL, e.g., overfitting on the rehearsal samples or hindering the model from encoding fresh information. In order to address catastrophic forgetting wi thout overfitting on the rehearsal samples, we propose Augmentation Stability Re hearsal (ASR) in this paper, which selects the most representative and discrimin ative samples by estimating the augmentation stability for rehearsal. Meanwhile, we design a matching strategy for ASR to dynamically update the rehearsal buffe r. In addition, we further propose Contrastive Continuity on Augmentation Stabil ity Rehearsal (C2ASR) based on ASR. We show that C2ASR is an upper bound of the Information Bottleneck (IB) principle, which suggests that C2ASR essentially pre serves as much information shared among seen task streams as possible to prevent catastrophic forgetting and dismisses the redundant information between previou s task streams and current task stream to free up the ability to encode fresh in formation. Our method obtains a great achievement compared with state-of-the-art CSSL methods on a variety of CSSL benchmarks.
****
Treating Pseudo-labels Generation as Image Matting for Weakly Supervised Semanti c Segmentation

Changwei Wang, Rongtao Xu, Shibiao Xu, Weiliang Meng, Xiaopeng Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 755-765
Generating accurate pseudo-labels under the supervision of image categories is a crucial step in Weakly Supervised Semantic Segmentation (WSSS). In this work, we propose a Mat-Label pipeline that provides a fresh way to treat WSSS pseudo-labels generation as an image matting task. By taking a trimap as input which specifies the foreground, background and unknown regions, the image matting task outputs an object mask with fine edges. The intuition behind our Mat-Label is that generating trimap is much easier than generating pseudo-labels directly under weakly supervised setting. Although current CAM-based methods are off-the-shelf solutions for generating a trimap, they suffer from cross-category and foreground-background pixel prediction confusion. To solve this problem, we develop a Double Decoupled Class Activation Map (D2CAM) for Mat-Label to generate a high-quality trimap. By drawing on the idea of metric learning, we explicitly model class activation map with category decoupling and foreground-background decoupling. We also design two simple yet effective refinement constraints for D2CAM to stabilize optimization and eliminate non-exclusive activation. Extensive experiments validate that our Mat-Label achieves substantial and consistent performance gains compared to current state-of-the-art WSSS approaches. Our code is available at supplementary material.
****
Structural Alignment for Network Pruning through Partial Regularization
Shangqian Gao, Zeyu Zhang, Yanfu Zhang, Feihu Huang, Heng Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17402-17412
In this paper, we propose a novel channel pruning method to reduce the computational and storage costs of Convolutional Neural Networks (CNNs). Many existing one-shot pruning methods directly remove redundant structures, which brings a huge gap between the model before and after network pruning. This gap will no doubt result in performance loss for network pruning. To mitigate this gap, we first learn a target sub-network during the model training process, and then we use this sub-network to guide the learning of model weights through partial regularization. The target sub-network is learned and produced by using an architecture generator, and it can be optimized efficiently. In addition, we also derive the proximal gradient for our proposed partial regularization to facilitate the structural alignment process. With these designs, the gap between the pruned model and the sub-network is reduced, thus improving the pruning performance. Empirical results also suggest that the sub-network found by our method has a much higher performance than the one-shot pruning setting. Extensive experiments show that our method can achieve state-of-the-art performances on CIFAR-10 and ImageNet with ResNets and MobileNet-V2.
****
Learning Long-Range Information with Dual-Scale Transformers for Indoor Scene Completion
Ziqi Wang, Fei Luo, Xiaoxiao Long, Wenxiao Zhang, Chunxia Xiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18569-18579
Due to the limited resolution of 3D sensors and the inevitable mutual occlusion between objects, 3D scans of real scenes are commonly incomplete.
Previous scene completion methods struggle to capture long-range spatial feature, resulting in unsatisfactory completion results.
To alleviate the problem, we propose a novel Dual-Scale Transformer Network (DST-Net) that efficiently utilizes both long-range and short-range spatial context information to improve the quality of 3D scene completion.
To reduce the heavy computation cost of extracting long-range features via transformers, DST-Net adopts a self-supervised two-stage completion strategy. In the first stage, we split the input scene into blocks, and perform completion on individual blocks. In the second stage, the blocks are merged together as a whole and then further refined to improve completeness.

More importantly, we propose a contrastive attention training strategy to encourage the transformers to learn distinguishable features for better scene completion. Experiments on datasets of Matterport3D, ScanNet, and ICL-NUIM demonstrate that our method can generate better completion results, and our method outperforms the state-of-the-art methods quantitatively and qualitatively.
****
A Game of Bundle Adjustment - Learning Efficient Convergence
Amir Belder, Refael Vivanti, Ayellet Tal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8428-8437
Bundle adjustment is the common way to solve localization and mapping. It is an iterative process in which a system of non-linear equations is solved using two optimization methods, weighted by a damping factor. In the classic approach, the latter is chosen heuristically by the Levenberg-Marquardt algorithm on each iteration. This might take many iterations, making the process computationally expensive, which might be harmful to real-time applications. We propose to replace this heuristic by viewing the problem in a holistic manner, as a game, and formulating it as a reinforcement-learning task. We set an environment which solves the non-linear equations and train an agent to choose the damping factor in a learned manner. We demonstrate that our approach considerably reduces the number of iterations required to reach the bundle adjustment's convergence, on both synthetic and real-life scenarios. We show that this reduction benefits the classic approach and can be integrated with other bundle adjustment acceleration methods.
****
Learning Correction Filter via Degradation-Adaptive Regression for Blind Single Image Super-Resolution
Hongyang Zhou, Xiaobin Zhu, Jianqing Zhu, Zheng Han, Shi-Xue Zhang, Jingyan Qin, Xu-Cheng Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12365-12375
Although existing image deep learning super-resolution (SR) methods achieve promising performance on benchmark datasets, they still suffer from severe performance drops when the degradation of the low-resolution (LR) input is not covered in training. To address the problem, we propose an innovative unsupervised method of Learning Correction Filter via Degradation-Adaptive Regression for Blind Single Image Super-Resolution. Highly inspired by the generalized sampling theory, our method aims to enhance the strength of off-the-shelf SR methods trained on known degradations and adapt to unknown complex degradations to generate improved results. Specifically, we first conduct degradation estimation for each local image region by learning the internal distribution in an unsupervised manner via GAN. Instead of assuming degradation are spatially invariant across the whole image, we learn correction filters to adjust degradations to known degradations in a spatially variant way by a novel linearly-assembled pixel degradation-adaptive regression module (DARM). DARM is lightweight and easy to optimize on a dictionary of multiple pre-defined filter bases. Extensive experiments on synthetic and real-world datasets verify the effectiveness of our method both qualitatively and quantitatively. Code can be available at: https://github.com/edbca/DARSR.
****
UMFuse: Unified Multi View Fusion for Human Editing Applications
Rishabh Jain, Mayur Hemani, Duygu Ceylan, Krishna Kumar Singh, Jingwan Lu, Mausoom Sarkar, Balaji Krishnamurthy; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7182-7191
Numerous pose-guided human editing methods have been explored by the vision community due to their extensive practical applications. However, most of these methods still use an image-to-image formulation in which a single image is given as input to produce an edited image as output. This objective becomes ill-defined in cases when the target pose differs significantly from the input pose. Existing methods then resort to in-painting or style transfer to handle occlusions and preserve content. In this paper, we explore the utilization of multiple views to minimize the issue of missing information and generate an accurate representatio

n
 of the underlying human model. To fuse knowledge from multiple viewpoints, we design a multi-view fusion network that takes the pose key points and texture from
 multiple source images and generates an explainable per pixel appearance retrieval map. Thereafter, the encodings from a separate network (trained on a single-view human reposing task) are merged in the latent space. This enables us to generate accurate, precise, and visually coherent images for different editing tasks. We show the application of our network on two newly proposed tasks - Multi-view human
 reposing and Mix&Match Human Image generation. Additionally, we study the limitations of single-view editing and scenarios in which multi-view provides a better alternative.
****
CROSSFIRE: Camera Relocalization On Self-Supervised Features from an Implicit Representation
Arthur Moreau, Nathan Piasco, Moussab Bennehar, Dzmitry Tsishkou, Bogdan Stanciulescu, Arnaud de La Fortelle; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 252-262
Beyond novel view synthesis, Neural Radiance Fields are useful for applications that interact with the real world. In this paper, we use them as an implicit map of a given scene and propose a camera relocalization algorithm tailored for this representation. The proposed method enables to compute in real-time the precise position of a device using a single RGB camera, during its navigation. In contrast with previous work, we do not rely on pose regression or photometric alignment but rather use dense local features obtained through volumetric rendering which are specialized on the scene with a self-supervised objective. As a result, our algorithm is more accurate than competitors, able to operate in dynamic outdoor environments with changing lightning conditions and can be readily integrated in any volumetric neural renderer.
****
Discriminative Class Tokens for Text-to-Image Diffusion Models
Idan Schwartz, Vésteinn Snæbjarnarson, Hila Chefer, Serge Belongie, Lior Wolf, Sagie Benaim; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22725-22735
Recent advances in text-to-image diffusion models have enabled the generation of diverse and high-quality images. While impressive, the images often fall short of depicting subtle details and are susceptible to errors due to ambiguity in the input text. One way of alleviating these issues is to train diffusion models on class-labeled datasets. This approach has two disadvantages: (i) supervised datasets are generally small compared to large-scale scraped text-image datasets on which text-to-image models are trained, affecting the quality and diversity of the generated images, or (ii) the input is a hard-coded label, as opposed to free-form text, limiting the control over the generated images.

 In this work, we propose a non-invasive fine-tuning technique that capitalizes on the expressive potential of free-form text while achieving high accuracy through discriminative signals from a pretrained classifier. This is done by iteratively modifying the embedding of an added input token of a text-to-image diffusion model, by steering generated images toward a given target class according to a classifier. Our method is fast compared to prior fine-tuning methods and does not require a collection of in-class images or retraining of a noise-tolerant classifier. We evaluate our method extensively, showing that the generated images are: (i) more accurate and of higher quality than standard diffusion models, (ii) can be used to augment training data in a low-resource setting, and (iii) reveal information about the data used to train the guiding classifier. The code is available at https://github.com/idansc/discriminative_class_tokens.
****
SINC: Spatial Composition of 3D Human Motions for Simultaneous Action Generation
Nikos Athanasiou, Mathis Petrovich, Michael J. Black, Gül Varol; Proceedings of

the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9984-9995
Our goal is to synthesize 3D human motions given textual inputs describing simultaneous actions, for example `waving hand' while `walking' at the same time. We refer to generating such simultaneous movements as performing `spatial compositions'. In contrast to `temporal compositions' that seek to transition from one action to another, spatial compositing requires understanding which body parts are involved with which action, to be able to move them simultaneously. Motivated by the observation that the correspondence between actions and body parts is encoded in powerful language models, we extract this knowledge by prompting GPT-3 with text such as "what are the body parts involved in the action <action name>?", while also providing the parts list and a few examples. Given this action-part mapping, we combine body parts from two motions together and establish the first automated method to spatially compose two actions. However, training data with compositional actions is always limited by the combinatorics. Hence, we further create synthetic data with this approach, and use it to train a new state-of-the-art text-to-motion generation model, called SINC ("SImultaneous actioN Compositions for 3D human motions"). In our experiments, we find that training with such GPT-guided synthetic data improves spatial composition generation over baselines.
Our code is publicly available at https://sinc.is.tue.mpg.de/.
****
ORC: Network Group-based Knowledge Distillation using Online Role Change
Junyong Choi, Hyeon Cho, Seokhwa Cheung, Wonjun Hwang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17381-17390
In knowledge distillation, since a single, omnipotent teacher network cannot solve all problems, multiple teacher-based knowledge distillations have been studied recently. However, sometimes their improvements are not as good as expected because some immature teachers may transfer the false knowledge to the student. In this paper, to overcome this limitation and take the efficacy of the multiple networks, we divide the multiple networks into teacher and student groups, respectively. That is, the student group is a set of immature networks that require learning the teacher's knowledge, while the teacher group consists of the selected networks that are capable of teaching successfully. We propose our online role change strategy where the top-ranked networks in the student group are able to promote to the teacher group at every iteration. After training the teacher group using the error samples of the student group to refine the teacher group's knowledge, we transfer the collaborative knowledge from the teacher group to the student group successfully. We verify the superiority of the proposed method on CIFAR-10, CIFAR-100, and ImageNet which achieves high performance. We further show the generality of our method with various backbone architectures such as ResNet, WRN, VGG, Mobilenet, and Shufflenet.
****
Audiovisual Masked Autoencoders
Mariana-Iuliana Georgescu, Eduardo Fonseca, Radu Tudor Ionescu, Mario Lucic, Cordelia Schmid, Anurag Arnab; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16144-16154
Can we leverage the audiovisual information already present in video to improve self-supervised representation learning? To answer this question, we study various pretraining architectures and objectives within the masked autoencoding framework, motivated by the success of similar methods in natural language and image understanding. We show that we can achieve significant improvements on audiovisual downstream classification tasks, surpassing the state-of-the-art on VGGSound and AudioSet. Furthermore, we can leverage our audiovisual pretraining scheme for multiple unimodal downstream tasks using a single audiovisual pretrained model. We additionally demonstrate the transferability of our representations, achieving state-of-the-art audiovisual results on Epic Kitchens without pretraining specifically for this dataset.
****
MV-DeepSDF: Implicit Modeling with Multi-Sweep Point Clouds for 3D Vehicle Recon

struction in Autonomous Driving
Yibo Liu, Kelly Zhu, Guile Wu, Yuan Ren, Bingbing Liu, Yang Liu, Jinjun Shan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8306-8316
Reconstructing 3D vehicles from noisy and sparse partial point clouds is of great significance to autonomous driving. Most existing 3D reconstruction methods cannot be directly applied to this problem because they are elaborately designed to deal with dense inputs with trivial noise. In this work, we propose a novel framework, dubbed MV-DeepSDF, which estimates the optimal Signed Distance Function (SDF) shape representation from multi-sweep point clouds
to reconstruct vehicles in the wild. Although there have been some SDF-based implicit modeling methods, they only focus on single-view-based reconstruction, resulting in low fidelity. In contrast, we first analyze multi-sweep consistency and complementarity in the latent feature space and propose to transform the implicit space shape estimation problem into an element-to-set feature extraction problem. Then, we devise a new architecture to extract individual element-level representations and aggregate them to generate a set-level predicted latent code. This set-level latent code is an expression of the optimal 3D shape in the implicit space, and can be subsequently decoded to a continuous SDF of the vehicle. In this way, our approach learns consistent and complementary information among multi-sweeps for 3D vehicle reconstruction. We conduct thorough experiments on two real-world autonomous driving datasets (Waymo and KITTI) to demonstrate the superiority of our approach over state-of-the-art alternative methods both qualitatively and quantitatively.
****
CHORD: Category-level Hand-held Object Reconstruction via Shape Deformation
Kailin Li, Lixin Yang, Haoyu Zhen, Zenan Lin, Xinyu Zhan, Licheng Zhong, Jian Xu, Kejian Wu, Cewu Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9444-9454
In daily life, humans utilize hands to manipulate objects. Modeling the shape of objects that are manipulated by the hand is essential for AI to comprehend daily tasks and to learn manipulation skills. However, previous approaches have encountered difficulties in reconstructing the precise shapes of hand-held objects, primarily owing to a deficiency in prior shape knowledge and inadequate data for training. As illustrated, given a particular type of tool, such as a mug, despite its infinite variations in shape and appearance, humans have a limited number of 'effective' modes and poses for its manipulation. This can be attributed to the fact that humans have mastered the shape prior of the 'mug' category, and can quickly establish the corresponding relations between different mug instances and the prior, such as where the rim and handle are located. In light of this, we propose a new method, CHORD, for Category-level Hand-held Object Reconstruction via shape Deformation. CHORD deforms a categorical shape prior for reconstructing the intra-class objects. To ensure accurate reconstruction, we empower CHORD with three types of awareness: appearance, shape, and interacting pose. In addition, we have constructed a new dataset, COMIC, of category-level hand-object interaction. COMIC contains a rich array of object instances, materials, hand interactions, and viewing directions. Extensive evaluation shows that CHORD outperforms state-of-the-art approaches in both quantitative and qualitative measures. Code, model, and datasets are available at https://kailinli.github.io/CHORD.
****
Unmasking Anomalies in Road-Scene Segmentation
Shyam Nandan Rai, Fabio Cermelli, Dario Fontanel, Carlo Masone, Barbara Caputo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4037-4046
Anomaly segmentation is a critical task for driving applications, and it is approached traditionally as a per-pixel classification problem. However, reasoning individually about each pixel without considering their contextual semantics results in high uncertainty around the objects' boundaries and numerous false positives. We propose a paradigm change by shifting from a per-pixel classification to a mask classification. Our mask-based method, Mask2Anomaly, demonstrates the fe

asibility of integrating an anomaly detection method in a mask-classification ar chitecture. Mask2Anomaly includes several technical novelties that are designed to improve the detection of anomalies in masks: i) a global masked attention mod ule to focus individually on the foreground and background regions; ii) a mask c ontrastive learning that maximizes the margin between an anomaly and known class es; and iii) a mask refinement solution to reduce false positives. Mask2Anomaly achieves new state-of-the-art results across a range of benchmarks, both in the per-pixel and component-level evaluations. In particular, Mask2Anomaly reduces t he average false positives rate by 60% wrt the previous state-of-the-art.
****
DomainDrop: Suppressing Domain-Sensitive Channels for Domain Generalization
Jintao Guo, Lei Qi, Yinghuan Shi; Proceedings of the IEEE/CVF International Conf erence on Computer Vision (ICCV), 2023, pp. 19114-19124
Deep Neural Networks have exhibited considerable success in various visual tasks . However, when applied to unseen test datasets, state-of-the-art models often s uffer performance degradation due to domain shifts. In this paper, we introduce a novel approach for domain generalization from a novel perspective of enhancing the robustness of channels in feature maps to domain shifts. We observe that mo dels trained on source domains contain a substantial number of channels that exh ibit unstable activations across different domains, which are inclined to captur e domain-specific features and behave abnormally when exposed to unseen target d omains. To address the issue, we propose a DomainDrop framework to continuously enhance the channel robustness to domain shifts, where a domain discriminator is used to identify and drop unstable channels in feature maps of each network lay er during forward propagation. We theoretically prove that our framework could e ffectively lower the generalization bound. Extensive experiments on several benc hmarks indicate that our framework achieves state-of-the-art performance compare d to other competing methods. Our code is available at https://github.com/linger inglight/DomainDrop.
****
Towards Universal LiDAR-Based 3D Object Detection by Multi-Domain Knowledge Tran sfer
Guile Wu, Tongtong Cao, Bingbing Liu, Xingxin Chen, Yuan Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8669-867 8
Contemporary LiDAR-based 3D object detection methods mostly focus on single-doma in learning or cross-domain adaptive learning. However, for autonomous driving s ystems, optimizing a specific LiDAR-based 3D object detector for each domain is costly and lacks of scalability in real-world deployment. It is desirable to tra in a universal LiDAR-based 3D object detector from multiple domains. In this wor k, we propose the first attempt to explore multi-domain learning and generalizat ion for LiDAR-based 3D object detection. We show that jointly optimizing a 3D ob ject detector from multiple domains achieves better generalization capability co mpared to the conventional single-domain learning model. To explore informative knowledge across domains towards a universal 3D object detector, we propose a mu lti-domain knowledge transfer framework with universal feature transformation. T his approach leverages spatial-wise and channel-wise knowledge across domains to learn universal feature representations, so it facilitates to optimize a univer sal 3D object detector for deployment at different domains. Extensive experiment s on four benchmark datasets (Waymo, KITTI, NuScenes and ONCE) show the superior ity of our approach over the state-of-the-art approaches for multi-domain learni ng and generalization in LiDAR-based 3D object detection.
****
StyleInV: A Temporal Style Modulated Inversion Network for Unconditional Video G eneration
Yuhan Wang, Liming Jiang, Chen Change Loy; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2023, pp. 22851-22861
Unconditional video generation is a challenging task that involves synthesizing high-quality videos that are both coherent and of extended duration. To address this challenge, researchers have used pretrained StyleGAN image generators for h

igh-quality frame synthesis and focused on motion generator design. The motion generator is trained in an autoregressive manner using heavy 3D convolutional discriminators to ensure motion coherence during video generation. In this paper, we introduce a novel motion generator design that uses a learning-based inversion network for GAN. The encoder in our method captures rich and smooth priors from encoding images to latents, and given the latent of an initially generated frame as guidance, our method can generate smooth future latent by modulating the inversion encoder temporally. Our method enjoys the advantage of sparse training and naturally constrains the generation space of our motion generator with the inversion network guided by the initial frame, eliminating the need for heavy discriminators. Moreover, our method supports style transfer with simple fine-tuning when the encoder is paired with a pretrained StyleGAN generator. Extensive experiments conducted on various benchmarks demonstrate the superiority of our method in generating long and high-resolution videos with decent single-frame quality and temporal consistency. Code is available at https://github.com/johannwyh/StyleInV.
****
Self-Calibrated Cross Attention Network for Few-Shot Segmentation
Qianxiong Xu, Wenting Zhao, Guosheng Lin, Cheng Long; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 655-665
The key to the success of few-shot segmentation (FSS) lies in how to effectively utilize support samples. Most solutions compress support foreground (FG) features into prototypes, but lose some spatial details. Instead, others use cross attention to fuse query features with uncompressed support FG. Query FG is safely fused with support FG, however, query background (BG) cannot find matched BG features in support FG, yet it inevitably integrates dissimilar features. Besides, as both query FG and BG are combined with support FG, they get entangled, thereby leading to ineffective segmentation. To cope with these issues, we design a self-calibrated cross attention (SCCA) block. For efficient patch-based attention, query and support features are firstly split into patches. Then, we design a patch alignment module to align each query patch with its most similar support patch for better cross attention. Specifically, SCCA takes a query patch as Q, and groups the patches from the same query image and the aligned patches from the support image as K&V. In this way, the query BG features are fused with matched BG features (from query patches), and thus the aforementioned issues will be mitigated. Moreover, when calculating SCCA, we design a scaled-cosine mechanism to better utilize the support features for similarity calculation. Extensive experiments conducted on PASCAL-5^i and COCO-20^i demonstrate the superiority of our model, e.g., the mIoU score under 5-shot setting on COCO-20^i is 5.6%+ better than previous state-of-the-arts. The code is available at https://github.com/Sam1224/SCCAN.
****
Anatomical Invariance Modeling and Semantic Alignment for Self-supervised Learning in 3D Medical Image Analysis
Yankai Jiang, Mingze Sun, Heng Guo, Xiaoyu Bai, Ke Yan, Le Lu, Minfeng Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15859-15869
Self-supervised learning (SSL) has recently achieved promising performance for 3D medical image analysis tasks. Most current methods follow existing SSL paradigm originally designed for photographic or natural images, which cannot explicitly and thoroughly exploit the intrinsic similar anatomical structures across varying medical images. This may in fact degrade the quality of learned deep representations by maximizing the similarity among features containing spatial misalignment information and different anatomical semantics. In this work, we propose a new self-supervised learning framework, namely Alice, that explicitly fulfills Anatomical invariance modeling and semantic alignment via elaborately combining discriminative and generative objectives. Alice introduces a new contrastive learning strategy which encourages the similarity between views that are diversely mined but with consistent high-level semantics, in order to learn invariant anatomical features. Moreover, we design a conditional anatomical feature alignment m

odule to complement corrupted embeddings with globally matched semantics and inter-patch topology information, conditioned by the distribution of local image content, which permits to create better contrastive pairs. Our extensive quantitative experiments on three 3D medical image analysis tasks demonstrate and validate the performance superiority of Alice, surpassing the previous best SSL counterpart methods and showing promising ability for united representation learning. Codes are available at https://github.com/alibaba-damo-academy/alice.
****
Towards High-Fidelity Text-Guided 3D Face Generation and Manipulation Using only Images
Cuican Yu, Guansong Lu, Yihan Zeng, Jian Sun, Xiaodan Liang, Huibin Li, Zongben Xu, Songcen Xu, Wei Zhang, Hang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15326-15337
Generating 3D faces from textual descriptions has a multitude of applications, such as gaming, movie and robotics. Recent progresses have demonstrated the success of unconditional 3D face generation and text-to-3D shape generation. However, due to the limited text-3D face data pairs, text-driven 3D face generation remains an open problem. In this paper, we propose a text-guided 3D faces generation method, refer as TG-3DFace, for generating realistic 3D face using text guidance. Specifically, we adopt an unconditional 3D face generation framework and equip it with text conditions, which learns the text-guided 3D face generation with only text-2D face data. On top of that, we propose two text-to-face cross-modal alignment techniques, including the global contrastive learning and the fine-grained alignment module, to facilitate high semantic consistency between generated 3D faces and input texts. Besides, we present directional classifier guidance during the inference process, which encourages creativity for out-of-domain generations. Compared to the existing methods, TG-3DFace creates more realistic and aesthetically pleasing 3D faces, boosting 9% multi-view consistency (MVIC) over Latent3D. The rendered face images generated by TG-3DFace achieve higher FID and CLIP score than text-to-2D face/image generation models, demonstrating our superiority in generating realistic and semantic-consistent textures.
****
SSDA: Secure Source-Free Domain Adaptation
Sabbir Ahmed, Abdullah Al Arafat, Mamshad Nayeem Rizve, Rahim Hossain, Zhishan Guo, Adnan Siraj Rakin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19180-19190
Source-free domain adaptation (SFDA) is a popular unsupervised domain adaptation method where a pre-trained model from a source domain is adapted to a target domain without accessing any source data. Despite rich results in this area, existing literature overlooks the security challenges of the unsupervised SFDA setting in presence of a malicious source domain owner. This work investigates the effect of a source adversary which may inject a hidden malicious behavior (Backdoor/Trojan) during source training and potentially transfer it to the target domain even after benign training by the victim (target domain owner). Our investigation of the current SFDA setting reveals that because of the unique challenges present in SFDA (e.g., no source data, target label), defending against backdoor attack using existing defenses become practically ineffective in protecting the target model. To address this, we propose a novel target domain protection scheme called secure source-free domain adaptation (SSDA). SSDA adopts a single-shot model compression of a pre-trained source model and a novel knowledge transfer scheme with a spectral-norm-based loss penalty for target training. The proposed static compression and the dynamic training loss penalty are designed to suppress the malicious channels responsive to the backdoor during the adaptation stage. At the same time, the knowledge transfer from an uncompressed auxiliary model helps to recover the benign test accuracy. Our extensive evaluation on multiple dataset and domain tasks against recent backdoor attacks reveal that the proposed SSDA can successfully defend against strong backdoor attacks with little to no degradation in test accuracy compared to the vulnerable baseline SFDA methods. Our code is available at https://github.com/ML-Security-Research-LAB/SSDA.
****

ENTL: Embodied Navigation Trajectory Learner
Klemen Kotar, Aaron Walsman, Roozbeh Mottaghi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10863-10872
We propose Embodied Navigation Trajectory Learner (ENTL), a method for extracting long sequence representations for embodied navigation. Our approach unifies world modeling, localization and imitation learning into a single sequence prediction task. We train our model using vector-quantized predictions of future states conditioned on current states and actions. ENTL's generic architecture enables the sharing of the the spatio-temporal sequence encoder for multiple challenging embodied tasks. We achieve competitive performance on navigation tasks using significantly less data than strong baselines while performing auxiliary tasks such as localization and future frame prediction (a proxy for world modeling). A key property of our approach is that the model is pre-trained without any explicit reward signal, which makes the resulting model generalizable to multiple tasks and environments.
****
AGG-Net: Attention Guided Gated-Convolutional Network for Depth Image Completion
Dongyue Chen, Tingxuan Huang, Zhimin Song, Shizhuo Deng, Tong Jia; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8853-8862
Recently, stereo vision based on lightweight RGBD cameras has been widely used in various fields. However, limited by the imaging principles, the commonly used RGB-D cameras based on TOF, structured light, or binocular vision acquire some invalid data inevitably, such as weak reflection, boundary shadows, and artifacts, which may bring adverse impacts to the follow-up work. In this paper, we propose a new model for depth image completion based on the Attention Guided Gated-convolutional Network (AGG-Net), through which more accurate and reliable depth images can be obtained based on the raw depth maps and the corresponding RGB images. Our model employs a UNet-like architecture which consists of two parallel branches of depth and color features. In the encoding stage, an Attention Guided Gated Convolution (AG-GConv) module is proposed to realize the fusion of depth and color features at different scales, which can effectively reduce the negative impacts of invalid depth data on the reconstruction. In the decoding stage, an Attention Guided Skip Connection (AG-SC) module is presented to avoid introducing too many depth-irrelevant features to the reconstruction. The experimental results demonstrate that our method outperforms the state-of-the-art methods on the popular benchmarks NYU-Depth V2, DIML, and SUN RGB-D.
****
Learning Global-aware Kernel for Image Harmonization
Xintian Shen, Jiangning Zhang, Jun Chen, Shipeng Bai, Yue Han, Yabiao Wang, Chengjie Wang, Yong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7535-7544
Image harmonization aims to solve the visual inconsistency problem in composited images by adaptively adjusting the foreground pixels with the background as references. Existing methods employ local color transformation or region matching between foreground and background, which neglects powerful proximity prior and independently distinguishes fore-/back-ground as a whole part for harmonization. As a result, they still show a limited performance across varied foreground objects and scenes. To address this issue, we propose a novel Global-aware Kernel Network (GKNet) to harmonize local regions with comprehensive consideration of long-distance background references.
Specifically, GKNet includes two parts, i.e., harmony kernel prediction and harmony kernel modulation branches. The former includes a Long-distance Reference Extractor (LRE) to obtain long-distance context and Kernel Prediction Blocks (KPB) to predict multi-level harmony kernels by fusing global information with local features. To achieve this goal, a novel Selective Correlation Fusion (SCF) module is proposed to better select relevant long-distance background references for local harmonization. The latter employs the predicted kernels to harmonize foreground regions with both local and global awareness. Abundant experiments demonstrate the superiority of our method for image harmonization over state-of-the-ar

t methods, e.g., achieving 39.53dB PSNR that surpasses the best counterpart by +0.78dB; decreasing fMSE by 11.5% and MSE by 6.7% compared with the SoTA method.
****

Real-Time Neural Rasterization for Large Scenes
Jeffrey Yunfan Liu, Yun Chen, Ze Yang, Jingkang Wang, Sivabalan Manivasagam, Raquel Urtasun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8416-8427
We propose a new method for realistic real-time novel-view synthesis (NVS) of large scenes. Existing fast neural rendering methods generate realistic results, but primarily work for small scale scenes (<50 square meter) and have difficulty at large scale (>10000 square meter). Traditional graphics-based rasterization rendering is fast for large scenes but lacks realism and requires expensive manually created assets. Our approach combines the best of both worlds by taking a moderate-quality scaffold mesh as input and learning a neural texture field and shader to model view-dependant effects to enhance realism, while still using the standard graphics pipeline for real-time rendering. Our method outperforms existing neural rendering methods, providing at least 30x faster rendering with comparable or better realism for large self-driving and drone scenes. Our work is the first to enable real-time visualization of large real-world scenes.
****

ESTextSpotter: Towards Better Scene Text Spotting with Explicit Synergy in Transformer
Mingxin Huang, Jiaxin Zhang, Dezhi Peng, Hao Lu, Can Huang, Yuliang Liu, Xiang Bai, Lianwen Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19495-19505
In recent years, end-to-end scene text spotting approaches are evolving to the Transformer-based framework. While previous studies have shown the crucial importance of the intrinsic synergy between text detection and recognition, recent advances in Transformer-based methods usually adopt an implicit synergy strategy with shared query, which can not fully realize the potential of these two interactive tasks. In this paper, we argue that the explicit synergy considering distinct characteristics of text detection and recognition can significantly improve the performance text spotting. To this end, we introduce a new model named Explicit Synergy-based Text Spotting Transformer framework (ESTextSpotter), which achieves explicit synergy by modeling discriminative and interactive features for text detection and recognition within a single decoder. Specifically, we decompose the conventional shared query into task-aware queries for text polygon and content, respectively. Through the decoder with the proposed vision-language communication module, the queries interact with each other in an explicit manner while preserving discriminative patterns of text detection and recognition, thus improving performance significantly. Additionally, we propose a task-aware query initialization scheme to ensure stable training. Experimental results demonstrate that our model significantly outperforms previous state-of-the-art methods. Code is available at https://github.com/mxin262/ESTextSpotter.
****

UGC: Unified GAN Compression for Efficient Image-to-Image Translation
Yuxi Ren, Jie Wu, Peng Zhang, Manlin Zhang, Xuefeng Xiao, Qian He, Rui Wang, Min Zheng, Xin Pan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17281-17291
Recent years have witnessed the prevailing progress of Generative Adversarial Networks (GANs) in image-to-image translation. However, the success of these GAN models hinges on ponderous computational costs and labor-expensive training data. Current efficient GAN learning techniques often fall into two orthogonal aspects: i) model slimming via reduced calculation costs; ii)data/label-efficient learning with fewer training data/labels. To combine the best of both worlds, we propose a new learning paradigm, Unified GAN Compression (UGC), with a unified optimization objective to seamlessly prompt the synergy of model-efficient and label-efficient learning. UGC sets up semi-supervised-driven network architecture search and adaptive online semi-supervised distillation stages sequentially, which formulates a heterogeneous mutual learning scheme to obtain an architecture-flex

ible, label-efficient, and performance-excellent model.
Extensive experiments demonstrate that UGC obtains state-of-the-art lightweight models even with less than 50% labels. UGC that compresses 40X MACs can achieve 21.43 FID on edges-shoes with 25% labels, which even outperforms the original model with 100% labels by 2.75 FID.
****
Efficient View Synthesis with Neural Radiance Distribution Field
Yushuang Wu, Xiao Li, Jinglu Wang, Xiaoguang Han, Shuguang Cui, Yan Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18506-18515
Recent work on Neural Radiance Fields (NeRF) has demonstrated significant advances in high-quality view synthesis. A major limitation of NeRF is its low rendering efficiency due to the need for multiple network forwardings to render a single pixel. Existing methods to improve NeRF either reduce the number of required samples or optimize the implementation to accelerate the network forwarding. Despite these efforts, the problem of multiple sampling persists due to the intrinsic representation of radiance fields. In contrast, Neural Light Fields (NeLF) reduce the computation cost of NeRF by querying only one single network forwarding per pixel. To achieve a close visual quality to NeRF, existing NeLF methods require significantly larger network capacities which limits their rendering efficiency in practice. In this work, we propose a new representation called Neural Radiance Distribution Field (NeRDF) that targets efficient view synthesis in real-time. Specifically, we use a small network similar to NeRF while preserving the rendering speed with a single network forwarding per pixel as in NeLF. The key is to model the radiance distribution along each ray with frequency basis and predict frequency weights using the network. Pixel values are then computed via volume rendering on radiance distributions. Experiments show that our proposed method offers a better trade-off among speed, quality, and network size than existing methods: we achieve a 254x speed-up over NeRF with similar network size, with only a marginal performance decline. Our project page is at yushuang-wu.github.io/NeRDF.
****
MixSpeech: Cross-Modality Self-Learning with Audio-Visual Stream Mixup for Visual Speech Translation and Recognition
Xize Cheng, Tao Jin, Rongjie Huang, Linjun Li, Wang Lin, Zehan Wang, Ye Wang, Huadai Liu, Aoxiong Yin, Zhou Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15735-15745
Multi-media communications facilitate global interaction among people. However, despite researchers exploring cross-lingual translation techniques such as machine translation and audio speech translation to overcome language barriers, there is still a shortage of cross-lingual studies on visual speech. This lack of research is mainly due to the absence of datasets containing visual speech and translated text pairs. In this paper, we present AVMuST-TED, the first dataset for Audio-Visual Multilingual Speech Translation, derived from TED talks. Nonetheless, visual speech is not as distinguishable as audio speech, making it difficult to develop a mapping from source speech phonemes to the target language text. To address this issue, we propose MixSpeech, a cross-modality self-learning framework that utilizes audio speech to regularize the training of visual speech tasks. To further minimize the cross-modality gap and its impact on knowledge transfer, we suggest adopting mixed speech, which is created by interpolating audio and visual streams, along with a curriculum learning strategy to adjust the mixing ratio as needed. MixSpeech enhances speech translation in noisy environments, improving BLEU scores for four languages on AVMuST-TED by +1.4 to +4.2. Moreover, it achieves state-of-the-art performance in lip reading on CMLR (11.1%), LRS2 (25.5%), and LRS3 (28.0%).
****
Chordal Averaging on Flag Manifolds and Its Applications
Nathan Mankovich, Tolga Birdal; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3881-3890
This paper presents a new, provably-convergent algorithm for computing the flag-

mean and flag-median of a set of points on a flag manifold under the chordal met
ric. The flag manifold is a mathematical space consisting of flags, which are se
quences of nested subspaces of a vector space that increase in dimension. The fl
ag manifold is a superset of a wide range of known matrix spaces, including Stie
fel and Grassmanians, making it a general object that is useful in a wide variet
y computer vision problems.

To tackle the challenge of computing first order flag statistics, we first tran
sform the problem into one that involves auxiliary variables constrained to the
Stiefel manifold. The Stiefel manifold is a space of orthogonal frames, and leve
raging the numerical stability and efficiency of Stiefel-manifold optimization e
nables us to compute the flag-mean effectively. Through a series of experiments,
we show the competence of our method in Grassmann and rotation averaging, as we
ll as principal component analysis.
****
Towards Building More Robust Models with Frequency Bias
Qingwen Bu, Dong Huang, Heming Cui; Proceedings of the IEEE/CVF International Co
nference on Computer Vision (ICCV), 2023, pp. 4402-4411
The vulnerability of deep neural networks to adversarial samples has been a majo
r impediment to their broad applications, despite their success in various field
s. Recently, some works suggested that adversarially-trained models emphasize th
e importance of low-frequency information to achieve higher robustness. While se
veral attempts have been made to leverage this frequency characteristic, they ha
ve all faced the issue that applying low-pass filters directly to input images l
eads to irreversible loss of discriminative information and poor generalizabilit
y to datasets with distinct frequency features. This paper presents a plug-and-p
lay module called the Frequency Preference Control Module that adaptively reconf
igures the low- and high-frequency components of intermediate feature representa
tions, providing better utilization of frequency in robust learning. Empirical s
tudies show that our proposed module can be easily incorporated into any adversa
rial training framework, further improving model robustness across different arc
hitectures and datasets. Additionally, experiments were conducted to examine how
the frequency bias of robust models impacts the adversarial training process an
d its final robustness, revealing interesting insights.
****
SparseBEV: High-Performance Sparse 3D Object Detection from Multi-Camera Videos
Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, Limin Wang; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18580-18590
Camera-based 3D object detection in BEV (Bird's Eye View) space has drawn great
attention over the past few years. Dense detectors typically follow a two-stage
pipeline by first constructing a dense BEV feature and then performing object de
tection in BEV space, which suffers from complex view transformations and high c
omputation cost. On the other side, sparse detectors follow a query-based paradi
gm without explicit dense BEV feature construction, but achieve worse performanc
e than the dense counterparts. In this paper, we find that the key to mitigate t
his performance gap is the adaptability of the detector in both BEV and image sp
ace. To achieve this goal, we propose SparseBEV, a fully sparse 3D object detect
or that outperforms the dense counterparts. SparseBEV contains three key designs
, which are (1) scale-adaptive self attention to aggregate features with adaptiv
e receptive field in BEV space, (2) adaptive spatio-temporal sampling to generat
e sampling locations under the guidance of queries, and (3) adaptive mixing to d
ecode the sampled features with dynamic weights from the queries. On the test sp
lit of nuScenes, SparseBEV achieves the state-of-the-art performance of 67.5 NDS
. On the val split, SparseBEV achieves 55.8 NDS while maintaining a real-time in
ference speed of 23.5 FPS. Code is available at https://github.com/MCG-NJU/Spars
eBEV.
****
Boosting Whole Slide Image Classification from the Perspectives of Distribution,
Correlation and Magnification
Linhao Qu, Zhiwei Yang, Minghong Duan, Yingfan Ma, Shuo Wang, Manning Wang, Zhij

ian Song; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21463-21473
Bag-based multiple instance learning (MIL) methods have become the mainstream for Whole Slide Image (WSI) classification. However, there are still three important issues that have not been fully addressed: (1) positive bags with a low positive instance ratio are prone to the influence of a large number of negative instances; (2) the correlation between local and global features of pathology images has not been fully modeled; and (3) there is a lack of effective information interaction between different magnifications. In this paper, we propose MILBooster, a powerful dual-scale multi-stage MIL framework to address these issues from the perspectives of distribution, correlation, and magnification. Specifically, to address issue (1), we propose a plug-and-play bag filter that effectively increases the positive instance ratio of positive bags. For issue (2), we propose a novel window-based Transformer architecture called PiceBlock to model the correlation between local and global features of pathology images. For issue (3), we propose a dual-branch architecture to process different magnifications and design an information interaction module called Scale Mixer for efficient information interaction between them. We conducted extensive experiments on four clinical WSI classification tasks using three datasets. MILBooster achieved new state-of-the-art performance on all these tasks. Codes will be available.
****
PolicyCleanse: Backdoor Detection and Mitigation for Competitive Reinforcement Learning
Junfeng Guo, Ang Li, Lixu Wang, Cong Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4699-4708
While real-world applications of reinforcement learning (RL) are becoming popular, the security and robustness of RL systems are worthy of more attention and exploration. In particular, recent works have revealed that, in a multi-agent RL environment, backdoor trigger actions can be injected into a victim agent (a.k.a. Trojan agent), which can result in a catastrophic failure as soon as it sees the backdoor trigger action. To ensure the security of RL agents against malicious backdoors, in this work, we propose the problem of Backdoor Detection in multi-agent RL systems, with the objective of detecting Trojan agents as well as the corresponding potential trigger actions, and further trying to mitigate their bad impact. In order to solve this problem, we propose PolicyCleanse that is based on the property that the activated Trojan agent's accumulated rewards degrade noticeably after several timesteps. Along with PolicyCleanse, we also design a machine unlearning-based approach that can effectively mitigate the detected backdoor. Extensive experiments demonstrate that
the proposed methods can accurately detect Trojan agents, and outperform existing backdoor mitigation baseline approaches by at least 3% in winning rate across various types of agents and environments.
****
Ref-NeuS: Ambiguity-Reduced Neural Implicit Surface Learning for Multi-View Reconstruction with Reflection
Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, Ying-Cong Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4251-4260
Neural implicit surface learning has shown significant progress in multi-view 3D reconstruction, where an object is represented by multilayer perceptrons that provide continuous implicit surface representation and view-dependent radiance. However, current methods often fail to accurately reconstruct reflective surfaces, leading to severe ambiguity. To overcome this issue, we propose Ref-NeuS, which aims to reduce ambiguity by attenuating the effect of reflective surfaces. Specifically, we utilize an anomaly detector to estimate an explicit reflection score with the guidance of multi-view context to localize reflective surfaces. Afterward, we design a reflection-aware photometric loss that adaptively reduces ambiguity by modeling rendered color as a Gaussian distribution, with the reflection score representing the variance. We show that together with a reflection direction-dependent radiance, our model achieves high-quality surface reconstruction on reflective surfaces and outperforms the state-of-the-arts by a large margin.

Besides, our model is also comparable on general surfaces.
****
Innovating Real Fisheye Image Correction with Dual Diffusion Architecture
Shangrong Yang, Chunyu Lin, Kang Liao, Yao Zhao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12699-12708
Fisheye image rectification is hindered by synthetic models producing poor results for real-world correction. To address this, we propose a Dual Diffusion Architecture (DDA) for fisheye rectification that offers better practicality. The DDA leverages Denoising Diffusion Probabilistic Models (DDPMs) to gradually introduce bidirectional noise, allowing the synthesized and real images to develop into a consistent noise distribution. As a result, our network can perceive the distribution of unlabelled real fisheye images without relying on a transfer network, thus improving the performance of real fisheye correction. Additionally, we design an unsupervised one-pass network that generates a plausible new condition to strengthen guidance and address the non-negligible indeterminacy between the prior condition and the target. It can significantly affect the rectification task, especially in cases where radial distortion causes significant artifacts. This network can be regarded as an alternate scheme for fast producing reliable results without iterative inference. Compared to the state-of-the-art methods, our approach achieves superior performance in both synthetic and real fisheye image corrections.
****
Global Perception Based Autoregressive Neural Processes
Jinyang Tai; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10487-10497
Increasingly, autoregressive approaches are being used to serialize observed variables based on specific criteria. The Neural Processes (NPs) model variable distribution as a continuous function and provide quick solutions for different tasks using a meta-learning framework. This paper proposes an autoregressive-based framework for NPs, based on their autoregressive properties. This framework leverages the autoregressive stacking effects of various variables to enhance the representation of the latent distribution, concurrently refining local and global relationships within the positional representation through the use of a sliding window
mechanism. Autoregression improves function approximations in a stacked fashion, thereby raising the upper bound of the optimization. We have designated this framework as Autoregressive Neural Processes (AENPs) or Conditional Autoregressive Neural Processes (CAENPs). Traditional NP models and their variants aim to capture relationships between the context sample points, without addressing either local or global considerations. Specifically, we capture contextual relationships in the deterministic path and introduce sliding window attention and global attention to reconcile local and global relationships in the context sample points. Autoregressive constraints exist between multiple latent variables in the latent paths, thus building a complex global structure that allows our model to learn complex
distributions. Finally, we demonstrate the effectiveness of the NPs or CFANPs models for 1D data, Bayesian optimization, and 2D data.
****
Class-incremental Continual Learning for Instance Segmentation with Image-level Weak Supervision
Yu-Hsing Hsieh, Guan-Sheng Chen, Shun-Xian Cai, Ting-Yun Wei, Huei-Fang Yang, Chu-Song Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1250-1261
Instance segmentation requires labor-intensive manual labeling of the contours of complex objects in images for training. The labels can also be provided incrementally in practice to balance the human labor in different time steps. However, research on incremental learning for instance segmentation with only weak labels is still lacking. In this paper, we propose a continual-learning method to segment object instances from image-level labels. Unlike most weakly-supervised instance segmentation (WSIS) which relies on traditional object proposals, we trans

fer the semantic knowledge from weakly-supervised semantic segmentation (WSSS) t
o WSIS to generate instance cues. To address the background shift problem in con
tinual learning, we employ the old class segmentation results generated by the p
revious model to provide more reliable semantic and peak hypotheses. To our know
ledge, this is the first work on weakly-supervised continual learning for instan
ce segmentation of images. Experimental results show that our method can achieve
 better performance on Pascal VOC and COCO datasets under various incremental se
ttings.
****
When Prompt-based Incremental Learning Does Not Meet Strong Pretraining
Yu-Ming Tang, Yi-Xing Peng, Wei-Shi Zheng; Proceedings of the IEEE/CVF Internati
onal Conference on Computer Vision (ICCV), 2023, pp. 1706-1716
Incremental learning aims to overcome catastrophic forgetting when learning deep
 networks from sequential tasks. With impressive learning efficiency and perform
ance, prompt-based methods adopt a fixed backbone to sequential tasks by learnin
g task-specific prompts. However, existing prompt-based methods heavily rely on
strong pretraining (typically trained on ImageNet-21k), and we find that their m
odels could be trapped if the potential gap between the pretraining task and unk
nown future tasks is large. In this work, we develop a learnable Adaptive Prompt
 Generator (APG). The key is to unify the prompt retrieval and prompt learning p
rocesses into a learnable prompt generator. Hence, the whole prompting process c
an be optimized to reduce the negative effects of the gap between tasks effectiv
ely. To make our APG avoid learning ineffective knowledge, we maintain a knowled
ge pool to regularize APG with the feature distribution of each class. Extensive
 experiments show that our method significantly outperforms advanced methods in
exemplar-free incremental learning without (strong) pretraining. Besides, under
strong retraining, our method also has comparable performance to existing prompt
-based models, showing that our method can still benefit from pretraining. Codes
 can be found at https://github.com/TOM-tym/APG
****
Multimodal High-order Relation Transformer for Scene Boundary Detection
Xi Wei, Zhangxiang Shi, Tianzhu Zhang, Xiaoyuan Yu, Lei Xiao; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22081-22
090
Scene boundary detection breaks down long videos into meaningful story-telling u
nits and plays a crucial role in high-level video understanding. Despite signifi
cant advancements in this area, this task remains a challenging problem as it re
quires a comprehensive understanding of multimodal cues and high-level semantics
. To tackle this issue, we propose a multimodal high-order relation transformer,
 which integrates a high-order encoder and an adaptive decoder in a unified fram
ework. By modeling the multimodal cues and exploring similarities between the sh
ots, the encoder is capable of capturing high-order relations between shots and
extracting shot features with context semantics. By clustering the shots adaptiv
ely, the decoder can discover more universal switch pattern between successive s
cenes, thus helping scene boundary detection. Extensive experimental results on
three standard benchmarks demonstrate that the proposed model performs favorably
 against state-of-the-art video scene detection methods.
****
Tri-MipRF: Tri-Mip Representation for Efficient Anti-Aliasing Neural Radiance Fi
elds
Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, Yuewen Ma; Proc
eedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023
, pp. 19774-19783
Despite the tremendous progress in neural radiance fields (NeRF), we still face
a dilemma of the trade-off between quality and efficiency, e.g., MipNeRF present
s fine-detailed and anti-aliased renderings but takes days for training, while I
nstant-ngp can accomplish the reconstruction in a few minutes but suffers from b
lurring or aliasing when rendering at various distances or resolutions due to ig
noring the sampling area. To this end, we propose a novel Tri-Mip encoding (a la
 "mipmap") that enables both instant reconstruction and anti-aliased high-fideli

ty rendering for neural radiance fields. The key is to factorize the pre-filtere d 3D feature spaces in three orthogonal mipmaps. In this way, we can efficiently perform 3D area sampling by taking advantage of 2D pre-filtered feature maps, w hich significantly elevates the rendering quality without sacrificing efficiency . To cope with the novel Tri-Mip representation, we propose a cone-casting rende ring technique to efficiently sample anti-aliased 3D features with the Tri-Mip e ncoding considering both pixel imaging and observing distance. Extensive experim ents on both synthetic and real-world datasets demonstrate our method achieves s tate-of-the-art rendering quality and reconstruction speed while maintaining a c ompact representation that reduces 25% model size compared against Instant-ngp. Code is available at the project webpage: https: //wbhu.github.io/projects/Tri-M ipRF
****
LaRS: A Diverse Panoptic Maritime Obstacle Detection Dataset and Benchmark
Lojze Žust, Janez Perš, Matej Kristan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20304-20314
The progress in maritime obstacle detection is hindered by the lack of a diverse dataset that adequately captures the complexity of general maritime environment s. We present the first maritime panoptic obstacle detection benchmark LaRS, fea turing scenes from Lakes, Rivers and Seas. Our major contribution is the new dat aset, which boasts the largest diversity in recording locations, scene types, ob stacle classes, and acquisition conditions among the related datasets. LaRS is c omposed of over 4000 per-pixel labeled key frames with nine preceding frames to allow utilization of the temporal texture, amounting to over 40k frames. Each ke y frame is annotated with 8 thing, 3 stuff classes and 19 global scene attribute s. We report the results of 27 semantic and panoptic segmentation methods, along with several performance insights and future research directions. To enable obj ective evaluation, we have implemented an online evaluation server. The LaRS dat aset, evaluation toolkit and benchmark are publicly available at: https://lojzez ust.github.io/lars-dataset
****
Exploring Transformers for Open-world Instance Segmentation
Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6611 -6621
Open-world instance segmentation is a rising task, which aims to segment all obj ects in the image by learning from a limited number of base-category objects. Th is task is challenging, as the number of unseen categories could be hundreds of times larger than that of seen categories. Recently, the DETR-like models have b een extensively studied in the closed world while stay unexplored in the open wo rld. In this paper, we utilize the Transformer for open-world instance segmentat ion and present SWORD. Firstly, we introduce to attach the stop-gradient operati on before classification head and further add IoU heads for discovering novel ob jects. We demonstrate that a simple stop-gradient operation not only prevents th e novel objects from being suppressed as background, but also allows the network to enjoy the merit of heuristic label assignment. Secondly, we propose a novel contrastive learning framework to enlarge the representations between objects an d background. Specifically, we maintain a universal object queue to obtain the o bject center, and dynamically select positive and negative samples from the obje ct queries for contrastive learning. While the previous works only focus on purs uing average recall and neglect average precision, we show the prominence of SWO RD by giving consideration to both criteria. Our models achieve state-of-the-art performance in various open-world cross-category and cross-dataset generalizati ons. Particularly, in VOC to non-VOC setup, our method sets new state-of-the-art results of 40.0% on ARb100 and 34.9% on ARm100. For COCO to UVO generalization, SWORD significantly outperforms the previous best open-world model by 5.9% on A Pm and 8.1% on ARm100, respectively.
****
VQA Therapy: Exploring Answer Differences by Visually Grounding Answers
Chongyan Chen, Samreen Anjum, Danna Gurari; Proceedings of the IEEE/CVF Internat

ional Conference on Computer Vision (ICCV), 2023, pp. 15315-15325
Visual question answering is a task of predicting the answer to a question about
 an image. Given that different people can provide different answers to a visual
 question, we aim to better understand why with answer groundings. We introduce
the first dataset that visually grounds each unique answer to each visual questi
on, which we call VQAAnswerTherapy. We then propose two novel problems of predic
ting whether a visual question has a single answer grounding and localizing all
answer groundings. We benchmark modern algorithms for these novel problems to sh
ow where they succeed and struggle. The dataset and evaluation server can be fou
nd publicly at https://vizwiz.org/tasks-and-datasets/vqa-answer-therapy/.
****
Energy-based Self-Training and Normalization for Unsupervised Domain Adaptation
Samitha Herath, Basura Fernando, Ehsan Abbasnejad, Munawar Hayat, Shahram Khadiv
i, Mehrtash Harandi, Hamid Rezatofighi, Gholamreza Haffari; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11653-1166
2
We propose an Unsupervised Domain Adaptation (UDA) method by making use of Energ
y-Based Learning (EBL) and demonstrate 1. EBL can be used to improve the instanc
e selection for a self-training task on the unlabelled target domain, and 2. ali
gnment and normalizing energy scores can learn domain-invariant representations.
 For the former, we show that an energy-based selection criterion can be used to
 model instance selections by mimicking the joint distribution between data and
predictions in the target domain. As per learning domain invariant representatio
ns, we show that stable domain alignment can be achieved by a combined energy al
ignment and an energy normalization process.
 We implement our method in consistent with the vision-transformer (ViT) backbon
e and empirically show that our proposed method can outperform state-of-the-art
ViT based UDA methods on diverse benchmarks (DomainNet, OfficeHome, and VISDA201
7).
****
Self-Evolved Dynamic Expansion Model for Task-Free Continual Learning
Fei Ye, Adrian G. Bors; Proceedings of the IEEE/CVF International Conference on
Computer Vision (ICCV), 2023, pp. 22102-22112
Task-Free Continual Learning (TFCL) aims to learn new concepts from a stream of
data without any task information. The Dynamic Expansion Model (DEM) has shown p
romising results in TFCL by dynamically expanding the model's capacity to deal w
ith shifts in the data distribution. However, existing approaches only consider
the recognition of the input shift as the expansion signal and ignore the correl
ation between the newly incoming data and previously learned knowledge, resultin
g in adding and training unnecessary parameters. In this paper, we propose a nov
el and effective framework for TFCL, which dynamically expands the architecture
of a DEM model through a self-assessment mechanism evaluating the diversity of k
nowledge among existing experts as expansion signals. This mechanism ensures lea
rning additional underlying data distributions with a compact model structure. A
 novelty-aware sample selection approach is proposed to manage the memory buffer
 that forces the newly added expert to learn novel information from a data strea
m, which further promotes the diversity among experts. Moreover, we also propose
 to reuse previously learned representation information for learning new incomin
g data by using knowledge transfer in TFCL, which has not been explored before.
The DEM expansion and training are regularized through a gradient updating mecha
nism to gradually explore the positive forward transfer, further improving the p
erformance. Empirical results on TFCL benchmarks show that the proposed framewor
k outperforms the state-of-the-art while using a reasonable number of parameters
. The code is available at https://github.com/dtuzi123/SEDEM/.
****
Adaptive Template Transformer for Mitochondria Segmentation in Electron Microsco
py Images
Yuwen Pan, Naisong Luo, Rui Sun, Meng Meng, Tianzhu Zhang, Zhiwei Xiong, Yongdon
g Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision
 (ICCV), 2023, pp. 21474-21484

Mitochondria, as tiny structures within the cell, are of significant importance
to study cell functions for biological and clinical analysis. And exploring how
to automatically segment mitochondria in electron microscopy (EM) images has att
racted increasing attention. However, most of existing methods struggle to adapt
 to different scales and appearances of the input due to the inherent limitation
s of the traditional CNN architecture. To mitigate these limitations, we propose
 a novel adaptive template transformer (ATFormer) for mitochondria segmentation.
 The proposed ATFormer model enjoys several merits. First, the designed structur
al template learning module can acquire appearance-adaptive templates of backgro
und, foreground and contour to sense the characteristics of different shapes of
mitochondria. And we further adopt an optimal transport algorithm to enlarge the
 discrepancy among diverse templates to fully activate corresponding regions. Se
cond, we introduce a hierarchical attention learning mechanism to absorb multi-l
evel information for templates to be adaptive scale-aware classifiers for dense
prediction. Extensive experimental results on three challenging benchmarks inclu
ding MitoEM, Lucchi and NucMM-Z datasets demonstrate that our ATFormer performs
favorably against state-of-the-art mitochondria segmentation methods.
****
BEVBert: Multimodal Map Pre-training for Language-guided Navigation
Dong An, Yuankai Qi, Yangguang Li, Yan Huang, Liang Wang, Tieniu Tan, Jing Shao;
 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
 2023, pp. 2737-2748
Large-scale pre-training has shown promising results on the vision-and-language
navigation (VLN) task. However, most existing pre-training methods employ discre
te panoramas to learn visual-textual associations. This requires the model to im
plicitly correlate incomplete, duplicate observations within the panoramas, whic
h may impair an agent's spatial understanding. Thus, we propose a new map-based
pre-training paradigm that is spatial-aware for use in VLN. Concretely, we build
 a local metric map to explicitly aggregate incomplete observations and remove d
uplicates, while modeling navigation dependency in a global topological map. Thi
s hybrid design can balance the demand of VLN for both short-term reasoning and
long-term planning. Then, based on the hybrid map, we devise a pre-training fram
ework to learn a multimodal map representation, which enhances spatial-aware cro
ss-modal reasoning thereby facilitating the language-guided navigation goal. Ext
ensive experiments demonstrate the effectiveness of the map-based pre-training r
oute for VLN, and the proposed method achieves state-of-the-art on four VLN benc
hmarks.
****
Collaborative Tracking Learning for Frame-Rate-Insensitive Multi-Object Tracking
Yiheng Liu, Junta Wu, Yi Fu; Proceedings of the IEEE/CVF International Conferenc
e on Computer Vision (ICCV), 2023, pp. 9964-9973
Multi-object tracking (MOT) at low frame rates can reduce computational, storage
 and power overhead to better meet the constraints of edge devices. Many existin
g MOT methods suffer from significant performance degradation in low-frame-rate
videos due to significant location and appearance changes between adjacent frame
s. To this end, we propose to explore collaborative tracking learning (CoTracker
) for frame-rate-insensitive MOT in a query-based end-to-end manner. Multiple hi
storical queries of the same target jointly track it with richer temporal descri
ptions. Meanwhile, we insert an information refinement module between every two
temporal blocking decoders to better fuse temporal clues and refine features. Mo
reover, a tracking object consistency loss is proposed to guide the interaction
between historical queries. Extensive experimental results demonstrate that in h
igh-frame-rate videos, CoTracker obtains higher performance than state-of-the-ar
t methods on large-scale datasets Dancetrack and BDD100K, and outperforms the ex
isting end-to-end methods on MOT17. More importantly, CoTracker has a significan
t advantage over state-of-the-art methods in low-frame-rate videos, which allows
 it to obtain faster processing speeds by reducing frame-rate requirements while
 maintaining higher performance. Code will be released at https://github.com/yol
omax/ColTrack
****

Tangent Model Composition for Ensembling and Continual Fine-tuning
Tian Yu Liu, Stefano Soatto; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18676-18686
Tangent Model Composition (TMC) is a method to combine component models independently fine-tuned around a pre-trained point. Component models are tangent vectors to the pre-trained model that can be added, scaled, or subtracted to support incremental learning, ensembling, or unlearning. Component models are composed at inference time via scalar combination, reducing the cost of ensembling to that of a single model. TMC improves accuracy by 4.2% compared to ensembling non-linearly fine-tuned models at a 2.5x to 10x reduction of inference cost, growing linearly with the number of component models. Each component model can be forgotten at zero cost, with no residual effect on the resulting inference. When used for continual fine-tuning, TMC is not constrained by sequential bias and can be executed in parallel on federated data. TMC outperforms recently published continual fine-tuning methods almost uniformly on each setting -- task-incremental, class-incremental, and data-incremental -- on a total of 13 experiments across 3 benchmark datasets, despite not using any replay buffer. TMC is designed for composing models that are local to a pre-trained embedding, but could be extended to more general settings.
****
Knowledge-Spreader: Learning Semi-Supervised Facial Action Dynamics by Consistifying Knowledge Granularity
Xiaotian Li, Xiang Zhang, Taoyue Wang, Lijun Yin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20979-20989
Recent studies on dynamic facial action unit (AU) detection have extensively relied on dense annotations. However, manual annotations are difficult, time-consuming, and costly. The canonical semi-supervised learning (SSL) methods ignore the consistency, extensibility, and adaptability of structural knowledge across spatial-temporal domains. Furthermore, the reliance on offline design and excessive parameters hinder the efficiency of the learning process. To remedy these issues, we propose a lightweight and on-line semi-supervised framework, a so-called Knowledge-Spreader (KS), to learn AU dynamics with sparse annotations. By formulating SSL as a Progressive Knowledge Distillation (PKD) problem, we aim to infer cross-domain information, specifically from spatial to temporal domains, by consistifying knowledge granularity within Teacher-Students Network. Specifically, KS employs sparsely annotated key-frames to learn AU dependencies as the privileged knowledge. Then, the model spreads the learned knowledge to their unlabeled neighbours by jointly applying knowledge distillation and pseudo-labeling, and completes the temporal information as the expanded knowledge. We term the progressive knowledge distillation as "Knowledge Spreading", which allows our model to learn spatial-temporal knowledge from video clips with only one label allocated. Extensive experiments demonstrate that KS achieves competitive performance as compared to the state of the arts under the circumstances of using only 2% labels on BP4D and 5% labels on DISFA. In addition, we have tested it on our newly developed large-scale comprehensive emotion database BP4D++, which contains considerable samples across well-synchronized and aligned sensor modalities for alleviating the scarcity issue of annotations and identities.
****
SSF: Accelerating Training of Spiking Neural Networks with Stabilized Spiking Flow
Jingtao Wang, Zengjie Song, Yuxi Wang, Jun Xiao, Yuran Yang, Shuqi Mei, Zhaoxiang Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5982-5991
Surrogate gradient (SG) is one of the most effective approaches for training spiking neural networks (SNNs). While assisting SNNs to achieve classification performance comparable to artificial neural networks, SG suffers from the problem of time-consuming training, preventing it from efficient learning. In this paper, we formally analyze the backward process of classic SG and find that the membrane accumulation through time leads to exponential growth of training time. With this discovery, we propose Stabilized Spiking Flow (SSF), a simple yet effective

approach to accelerate training of SG-based SNNs. For each spiking neuron, SSF a
verages its input and output activations over time to yield stabilized input and
 output, respectively. Then, instead of back propagating all errors that are rel
ated to current neuron and inherently entangled in time domain, the auxiliary gr
adient is directly propagated from the stabilized output to input through a devi
sed relationship mapping. Additionally, SSF method is suitable to different neur
on models. Extensive experiments on both static and neuromorphic datasets demons
trate that SNNs trained with SSF approach can achieve performance comparable to
the original counterparts, while reducing the training time significantly. In pa
rticular, SSF speeds up the training process of state-of-the-art SNN models up t
o 10x when time steps equal to 80.
****
Manipulate by Seeing: Creating Manipulation Controllers from Pre-Trained Represe
ntations
Jianren Wang, Sudeep Dasari, Mohan Kumar Srirama, Shubham Tulsiani, Abhinav Gupt
a; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV
), 2023, pp. 3859-3868
The field of visual representation learning has seen explosive growth in the pas
t years, but its benefits in robotics have been surprisingly limited so far. Pri
or work uses generic visual representations as a basis to learn (task-specific)
robot action policies (e.g., via behavior cloning). While the visual representat
ions do accelerate learning, they are primarily used to encode visual observatio
ns. Thus, action information has to be derived purely from robot data, which is
expensive to collect! In this work, we present a scalable alternative where the
visual representations can help directly infer robot actions. We observe that vi
sion encoders express relationships between image observations as distances (e.g
., via embedding dot product) that could be used to efficiently plan robot behav
ior. We operationalize this insight and develop a simple algorithm for acquiring
 a distance function and dynamics predictor, by fine-tuning a pre-trained repres
entation on human collected video sequences. The final method is able to substan
tially outperform traditional robot learning baselines (e.g., 70% success v.s. 5
0% for behavior cloning on pick-place) on a suite of diverse real-world manipula
tion tasks. It can also generalize to novel objects, without using any robot dem
onstrations during train time. For visualizations of the learned policies please
 check: https://agi-labs.github.io/manipulate-by-seeing/
****
Learning Human-Human Interactions in Images from Weak Textual Supervision
Morris Alper, Hadar Averbuch-Elor; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2023, pp. 2887-2899
Interactions between humans are diverse and context-dependent, but previous work
s have treated them as categorical, disregarding the heavy tail of possible inte
ractions. We propose a new paradigm of learning human-human interactions as free
 text from a single still image, allowing for flexibility in modeling the unlimi
ted space of situations and relationships between people. To overcome the absenc
e of data labelled specifically for this task, we use knowledge distillation app
lied to synthetic caption data produced by a large language model without explic
it supervision. We show that the pseudo-labels produced by this procedure can be
 used to train a captioning model to effectively understand human-human interact
ions in images, as measured by a variety of metrics that measure textual and sem
antic faithfulness and factual groundedness of our predictions. We further show
that our approach outperforms SOTA image captioning and situation recognition mo
dels on this task. We will release our code and pseudo-labels along with Waldo a
nd Wenda, a manually-curated test set for still image human-human interaction un
derstanding.
****
Prompt-aligned Gradient for Prompt Tuning
Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, Hanwang Zhang; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15659-15669
Thanks to the large pre-trained vision-language models (VLMs) like CLIP, we can
craft a zero-shot classifier by discrete prompt design, e.g., the confidence sco

re of an image being "[CLASS]" can be obtained by using the VLM provided similarity between the image and the prompt sentence "a photo of a [CLASS]". Furthermore, prompting shows great potential for fast adaptation of VLMs to downstream tasks if we fine-tune the soft prompts with few samples. However, we find a common failure that improper fine-tuning or learning with extremely few-shot samples may even under-perform the zero-shot prediction. Existing methods still address this problem by using traditional anti-overfitting techniques such as early stopping and data augmentation, which lack a principled solution specific to prompting. In this paper, we present Prompt-aligned Gradient, dubbed ProGrad to prevent prompt tuning from forgetting the general knowledge learned from VLMs. In particular, ProGrad only updates the prompt whose gradient is aligned (or non-conflicting) to the general knowledge, which is represented as the optimization direction offered by the predefined prompt predictions. Extensive experiments under the few-shot learning, domain generalization, base-to-new generalization and cross-dataset transfer settings demonstrate
the stronger few-shot generalization ability of ProGrad over state-of-the-art prompt tuning methods. Codes and theoretical proof are in Appendix.
****
Aperture Diffraction for Compact Snapshot Spectral Imaging
Tao Lv, Hao Ye, Quan Yuan, Zhan Shi, Yibo Wang, Shuming Wang, Xun Cao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10574-10584
We demonstrate a compact, cost-effective snapshot spectral imaging system named Aperture Diffraction Imaging Spectrometer (ADIS), which consists only of an imaging lens with an ultra-thin orthogonal aperture mask and a mosaic filter sensor, requiring no additional physical footprint compared to common RGB cameras. Then we introduce a new optical design that each point in the object space is multiplexed to discrete encoding locations on the mosaic filter sensor by diffraction-based spatial-spectral projection engineering generated from the orthogonal mask. The orthogonal projection is uniformly accepted to obtain a weakly calibration-dependent data form to enhance modulation robustness. Meanwhile, the Cascade Shift-Shuffle Spectral Transformer (CSST) with strong perception of the diffraction degeneration is designed to solve a sparsity-constrained inverse problem, realizing the volume reconstruction from 2D measurements with Large amount of aliasing. Our system is evaluated by elaborating the imaging optical theory and reconstruction algorithm with demonstrating the experimental imaging under a single exposure. Ultimately, we achieve the sub-super-pixel spatial resolution and high spectral resolution imaging. The code will be available at: https://github.com/Krito-ex/CSST.
****
Diffusion Action Segmentation
Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, Chang Xu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10139-10149
Temporal action segmentation is crucial for understanding long-form videos. Previous works on this task commonly adopt an iterative refinement paradigm by using multi-stage models. We propose a novel framework via denoising diffusion models, which nonetheless shares the same inherent spirit of such iterative refinement. In this framework, action predictions are iteratively generated from random noise with input video features as conditions. To enhance the modeling of three striking characteristics of human actions, including the position prior, the boundary ambiguity, and the relational dependency, we devise a unified masking strategy for the conditioning inputs in our framework. Extensive experiments on three benchmark datasets, i.e., GTEA, 50Salads, and Breakfast, are performed and the proposed method achieves superior or comparable results to state-of-the-art methods, showing the effectiveness of a generative approach for action segmentation.
****
Prototype Reminiscence and Augmented Asymmetric Knowledge Aggregation for Non-Exemplar Class-Incremental Learning
Wuxuan Shi, Mang Ye; Proceedings of the IEEE/CVF International Conference on Com

puter Vision (ICCV), 2023, pp. 1772-1781
Non-exemplar class-incremental learning (NECIL) requires deep models to maintain existing knowledge while continuously learning new classes without saving old class samples. In NECIL methods, prototypical representations are usually stored, which inject information from former classes to resist catastrophic forgetting in subsequent incremental learning. However, since the model continuously learns new knowledge, the stored prototypical representations cannot correctly model the properties of old classes in the existence of knowledge updates. To address this problem, we propose a novel prototype reminiscence mechanism that incorporates the previous class prototypes with arriving new class features to dynamically reshape old class feature distributions thus preserving the decision boundaries of previous tasks. In addition, to improve the model generalization on both newly arriving classes and old classes, we contribute an augmented asymmetric knowledge aggregation approach, which aggregates the overall knowledge of the current task and extracts the valuable knowledge of the past tasks, on top of self-supervised label augmentation. Experimental results on three benchmarks suggest the superior performance of our approach over the SOTA methods.
****
Exemplar-Free Continual Transformer with Convolutions
Anurag Roy, Vinay K. Verma, Sravan Voonna, Kripabandhu Ghosh, Saptarshi Ghosh, Abir Das; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5897-5907
Continual Learning (CL) involves training a machine learning model in a sequential manner to learn new information while retaining previously learned tasks without the presence of previous training data. Although there has been significant interest in CL, most recent CL approaches in computer vision have focused on convolutional architectures only. However, with the recent success of vision transformers, there is a need to explore their potential for CL. Although there have been some recent CL approaches for vision transformers, they either store training instances of previous tasks or require a task identifier during test time, which can be limiting. This paper proposes a new exemplar-free approach for class/task incremental learning called ConTraCon, which does not require task-id to be explicitly present during inference and avoids the need for storing previous training instances. The proposed approach leverages the transformer architecture and involves re-weighting the key, query, and value weights of the multi-head self-attention layers of a transformer trained on a similar task. The re-weighting is done using convolution, which enables the approach to maintain low parameter requirements per task. Additionally, an image augmentation-based entropic task identification approach is used to predict tasks without requiring task-ids during inference. Experiments on four benchmark datasets demonstrate that the proposed approach outperforms several competitive approaches while requiring fewer parameters.
****
Scalable Video Object Segmentation with Simplified Framework
Qiangqiang Wu, Tianyu Yang, Wei Wu, Antoni B. Chan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13879-13889
The current popular methods for video object segmentation (VOS) implement feature matching through several hand-crafted modules that separately perform feature extraction and matching. However, the above hand-crafted designs empirically cause insufficient target interaction, thus limiting the dynamic target-aware feature learning in VOS. To tackle these limitations, this paper presents a scalable Simplified VOS (SimVOS) framework to perform joint feature extraction and matching by leveraging a single transformer backbone. Specifically, SimVOS employs a scalable ViT backbone for simultaneous feature extraction and matching between query and reference features. This design enables SimVOS to learn better target-aware features for accurate mask prediction. More importantly, SimVOS could directly apply well-pretrained ViT backbones (e.g., MAE) for VOS, which bridges the gap between VOS and large-scale self-supervised pre-training. To achieve a better performance-speed trade-off, we further explore within-frame attention and propose a new token refinement module to improve the running speed and save computatio

nal cost. Experimentally, our SimVOS achieves state-of-the-art results on popula
r video object segmentation benchmarks, i.e., DAVIS-2017 (88.0% J&F), DAVIS-2016
 (92.9% J&F) and YouTube-VOS 2019 (84.2% J&F), without applying any synthetic vi
deo or BL30K pre-training used in previous VOS approaches. Our code and models a
re available at https://github.com/jimmy-dq/SimVOS.git.
****
Rehearsal-Free Domain Continual Face Anti-Spoofing: Generalize More and Forget L
ess
Rizhao Cai, Yawen Cui, Zhi Li, Zitong Yu, Haoliang Li, Yongjian Hu, Alex Kot; Pr
oceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 20
23, pp. 8037-8048
Face Anti-Spoofing (FAS) is recently studied under the continual learning settin
g, where the FAS models are expected to evolve after encountering data from new
domains. However, existing methods need extra replay buffers to store previous d
ata for rehearsal, which becomes infeasible when previous data is unavailable be
cause of privacy issues. In this paper, we propose the first rehearsal-free meth
od for Domain Continual Learning (DCL) of FAS, which deals with catastrophic for
getting and unseen domain generalization problems simultaneously. For better gen
eralization to unseen domains, we design the Dynamic Central Difference Convolut
ional Adapter (DCDCA) to adapt Vision Transformer (ViT) models during the contin
ual learning sessions. To alleviate the forgetting of previous domains without u
sing previous data, we propose the Proxy Prototype Contrastive Regularization (P
PCR) to constrain the continual learning with previous domain knowledge from the
 proxy prototypes. Simulating practical DCL scenarios, we devise two new protoco
ls which evaluate both generalization and anti-forgetting performance. Extensive
 experimental results show that our proposed method can improve the generalizati
on performance in unseen domains and alleviate the catastrophic forgetting of pr
evious knowledge. The code and protocol files are released on https://github.com
/RizhaoCai/DCL-FAS-ICCV2023.
****
Efficient Decision-based Black-box Patch Attacks on Video Recognition
Kaixun Jiang, Zhaoyu Chen, Hao Huang, Jiafeng Wang, Dingkang Yang, Bo Li, Yan Wa
ng, Wenqiang Zhang; Proceedings of the IEEE/CVF International Conference on Comp
uter Vision (ICCV), 2023, pp. 4379-4389
Although Deep Neural Networks (DNNs) have demonstrated excellent performance, th
ey are vulnerable to adversarial patches that introduce perceptible and localize
d perturbations to the input. Generating adversarial patches on images has recei
ved much attention, while adversarial patches on videos have not been well inves
tigated. Further, decision-based attacks, where attackers only access the predic
ted hard labels by querying threat models, have not been well explored on video
models either, even if they are practical in real-world video recognition scenes
. The absence of such studies leads to a huge gap in the robustness assessment f
or video models. To bridge this gap, this work first explores decision-based pat
ch attacks on video models. We analyze that the huge parameter space brought by
videos and the minimal information returned by decision-based models both greatl
y increase the attack difficulty and query burden. To achieve a query-efficient
attack, we propose a spatial-temporal differential evolution (STDE) framework. F
irst, STDE introduces target videos as patch textures and only adds patches on k
eyframes that are adaptively selected by temporal difference. Second, STDE takes
 minimizing the patch area as the optimization objective and adopts spatial-temp
oral mutation and crossover to search for the global optimum without falling int
o the local optimum. Experiments show STDE has demonstrated state-of-the-art per
formance in terms of threat, efficiency and imperceptibility. Hence, STDE has th
e potential to be a powerful tool for evaluating the robustness of video recogni
tion models.
****
Kick Back & Relax: Learning to Reconstruct the World by Watching SlowTV
Jaime Spencer, Chris Russell, Simon Hadfield, Richard Bowden; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15768-15
779

Self-supervised monocular depth estimation (SS-MDE) has the potential to scale to vast quantities of data. Unfortunately, existing approaches limit themselves to the automotive domain, resulting in models incapable of generalizing to complex environments such as natural or indoor settings. To address this, we propose a large-scale SlowTV dataset curated from YouTube, containing an order of magnitude more data than existing automotive datasets. SlowTV contains 1.7M images from a rich diversity of environments, such as worldwide seasonal hiking, scenic driving and scuba diving. Using this dataset, we train an SS-MDE model that provides zero-shot generalization to a large collection of indoor/outdoor datasets. The resulting model outperforms all existing SSL approaches and closes the gap on supervised SoTA, despite using a more efficient architecture. We additionally introduce a collection of best-practices to further maximize performance and zero-shot generalization. This includes 1) aspect ratio augmentation, 2) camera intrinsic estimation, 3) support frame randomization and 4) flexible motion estimation.

****

MetaGCD: Learning to Continually Learn in Generalized Category Discovery

Yanan Wu, Zhixiang Chi, Yang Wang, Songhe Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1655-1665

In this paper, we consider a real-world scenario where a model that is trained on pre-defined classes continually encounters unlabeled data that contains both known and novel classes. The goal is to continually discover novel classes while maintaining the performance in known classes. We name the setting Continual Generalized Category Discovery (C-GCD). Existing methods for novel class discovery cannot directly handle the C-GCD setting due to some unrealistic assumptions, such as the unlabeled data only containing novel classes. Furthermore, they fail to discover novel classes in a continual fashion. In this work, we lift all these assumptions and propose an approach, called MetaGCD, to learn how to incrementally discover with less forgetting. Our proposed method uses a meta-learning framework and leverages the offline labeled data to simulate the testing incremental learning process. A meta-objective is defined to revolve around two conflicting learning objectives to achieve novel class discovery without forgetting. Furthermore, a soft neighborhood-based contrastive network is proposed to discriminate uncorrelated images while attracting correlated images. We build strong baselines and conduct extensive experiments on three widely used benchmarks to demonstrate the superiority of our method.

****

Strip-MLP: Efficient Token Interaction for Vision MLP

Guiping Cao, Shengda Luo, Wenjian Huang, Xiangyuan Lan, Dongmei Jiang, Yaowei Wang, Jianguo Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1494-1504

Token interaction operation is one of the core modules in MLP-based models to exchange and aggregate information between different spatial locations. However, the power of token interaction on the spatial dimension is highly dependent on the spatial resolution of the feature maps, which limits the model's expressive ability, especially in deep layers where the feature are down-sampled to a small spatial size. To address this issue, we present a novel method called Strip-MLP to enrich the token interaction power in three ways. Firstly, we introduce a new MLP paradigm called Strip MLP layer that allows the token to interact with other tokens in a cross-strip manner, enabling the tokens in a row (or column) contribute to the information aggregations in adjacent but different strips of rows (or columns). Secondly, a Cascade Group Strip Mixing Module (CGSMM) is proposed to overcome the performance degradation caused by small spatial feature size. The module allows tokens to interact more effectively in the manners of within-patch and cross-patch, which is independent to the feature spatial size. Finally, based on the Strip MLP layer, we propose a novel Local Strip Mixing Module (LSMM) to boost the token interaction power in the local region. Extensive experiments demonstrate that Strip-MLP significantly improves the performance of MLP-based models on small datasets and obtains comparable or even better results on ImageNet with great superiorities on the number of parameters and FLOPs. In particular,

Strip-MLP models achieve higher average Top-1 accuracy than existing MLP-based models by +2.44% on Caltech-101 and +2.16% on CIFAR-100. The source codes will be available at https://github.com/Med-Process/Strip_MLP.
****
SAFARI: Versatile and Efficient Evaluations for Robustness of Interpretability
Wei Huang, Xingyu Zhao, Gaojie Jin, Xiaowei Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1988-1998
Interpretability of Deep Learning (DL) is a barrier to trustworthy AI. Despite great efforts made by the Explainable AI (XAI) community, explanations lack robustness--indistinguishable input perturbations may lead to different XAI results. Thus, it is vital to assess how robust DL interpretability is, given an XAI method. In this paper, we identify several challenges that the state-of-the-art is unable to cope with collectively: i) existing metrics are not comprehensive ii) XAI techniques are highly heterogeneous; iii) misinterpretations are normally rare events. To tackle these challenges, we introduce two black-box evaluation methods, concerning the worst-case interpretation discrepancy and a probabilistic notion of how robust in general, respectively. Genetic Algorithm (GA) with bespoke fitness function is used to solve constrained optimisation for efficient worst-case evaluation. Subset Simulation (SS), dedicated to estimating rare event probabilities, is used for evaluating overall robustness. Experiments show that the accuracy, sensitivity, and efficiency of our methods outperform the state-of-the-arts. Finally, we demonstrate two applications of our methods: ranking robust XAI methods and selecting training schemes to improve both classification and interpretation robustness.
****
ChildPlay: A New Benchmark for Understanding Children's Gaze Behaviour
Samy Tafasca, Anshul Gupta, Jean-Marc Odobez; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20935-20946
Gaze behaviors such as eye-contact or shared attention are important markers for diagnosing developmental disorders in children. While previous studies have looked at some of these elements, the analysis is usually performed on private datasets and is restricted to lab settings. Furthermore, all publicly available gaze target prediction benchmarks mostly contain instances of adults, which makes models trained on them less applicable to scenarios with young children. In this paper, we propose the first study for predicting the gaze target of children and interacting adults. To this end, we introduce the ChildPlay dataset: a curated collection of short video clips featuring children playing and interacting with adults in uncontrolled environments (e.g. kindergarten, therapy centers, preschools etc.), which we annotate with rich gaze information. We further propose a new model for gaze target prediction that is geometrically grounded by explicitly identifying the scene parts in the 3D field of view (3DFoV) of the person, leveraging recent geometry preserving depth inference methods. Our model achieves state of the art results on benchmark datasets and ChildPlay. Furthermore, results show that looking at faces prediction performance on children is much worse than on adults, and can be significantly improved by fine-tuning models using child gaze annotations. Our dataset is available at https://www.idiap.ch/en/dataset/childplay-gaze.
****
Towards General Low-Light Raw Noise Synthesis and Modeling
Feng Zhang, Bin Xu, Zhiqiang Li, Xinran Liu, Qingbo Lu, Changxin Gao, Nong Sang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10820-10830
Modeling and synthesizing low-light raw noise is a fundamental problem for computational photography and image processing applications. Although most recent works have adopted physics-based models to synthesize noise, the signal-independent noise in low-light conditions is far more complicated and varies dramatically across camera sensors, which is beyond the description of these models. To address this issue, we introduce a new perspective to synthesize the signal-independent noise by a generative model. Specifically, we synthesize the signal-dependent and signal-independent noise in a physics- and learning-based manner, respective

ly. In this way, our method can be considered as a general model, that is, it can simultaneously learn different noise characteristics for different ISO levels and generalize to various sensors. Subsequently, we present an effective multi-scale discriminator termed Fourier transformer discriminator (FTD) to distinguish the noise distribution accurately. Additionally, we collect a new low-light raw denoising (LRD) dataset for training and benchmarking. Qualitative validation shows that the noise generated by our proposed noise model can be highly similar to the real noise in terms of distribution. Furthermore, extensive denoising experiments demonstrate that our method performs favorably against state-of-the-art methods on different sensors.
****
Combating Noisy Labels with Sample Selection by Mining High-Discrepancy Examples
Xiaobo Xia, Bo Han, Yibing Zhan, Jun Yu, Mingming Gong, Chen Gong, Tongliang Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1833-1843
The sample selection approach is popular in learning with noisy labels. The state-of-the-art methods train two deep networks simultaneously for sample selection, which aims to employ their different learning abilities. To prevent two networks from converging to a consensus, their divergence should be maintained. Prior work presents that the divergence can be kept by locating the disagreement data on which the prediction labels of the two networks are different. However, this procedure is sample-inefficient for generalization, which means that only a few clean examples can be utilized in training. In this paper, to address the issue, we propose a simple yet effective method called CoDis. In particular, we select possibly clean data that simultaneously have high-discrepancy prediction probabilities between two networks. As selected data have high discrepancies in probabilities, the divergence of two networks can be maintained by training on such data. Additionally, the condition of high discrepancies is milder than disagreement, which allows more data to be considered for training, and makes our method more sample-efficient. Moreover, we show that the proposed method enables to mine hard clean examples to help generalization. Empirical results show that CoDis is superior to multiple baselines in the robustness of trained models.
****
Beyond the Pixel: a Photometrically Calibrated HDR Dataset for Luminance and Color Prediction
Christophe Bolduc, Justine Giroux, Marc Hébert, Claude Demers, Jean-François Lalonde; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8071-8081
Light plays an important role in human well-being. However, most computer vision tasks treat pixels without considering their relationship to physical luminance. To address this shortcoming, we introduce the Laval Photometric Indoor HDR Dataset, the first large-scale photometrically calibrated dataset of high dynamic range 360deg panoramas. Our key contribution is the calibration of an existing, uncalibrated HDR Dataset. We do so by accurately capturing RAW bracketed exposures simultaneously with a professional photometric measurement device (chroma meter) for multiple scenes across a variety of lighting conditions. Using the resulting measurements, we establish the calibration coefficients to be applied to the HDR images. The resulting dataset is a rich representation of indoor scenes which displays a wide range of illuminance and color, and varied types of light sources. We exploit the dataset to introduce three novel tasks, where: per-pixel luminance, per-pixel color and planar illuminance can be predicted from a single input image. Finally, we also capture another smaller photometric dataset with a commercial 360deg camera, to experiment on generalization across cameras. We are optimistic that the release of our datasets and associated code will spark interest in physically accurate light estimation within the community. Dataset and code are available at https://lvsn.github.io/beyondthepixel/.
****
What can Discriminator do? Towards Box-free Ownership Verification of Generative Adversarial Networks
Ziheng Huang, Boheng Li, Yan Cai, Run Wang, Shangwei Guo, Liming Fang, Jing Chen

, Lina Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5009-5019
In recent decades, Generative Adversarial Network (GAN) and its variants have achieved unprecedented success in image synthesis. However, well-trained GANs are under the threat of illegal steal or leakage. The prior studies on remote ownership verification assume a black-box setting where the defender can query the suspicious model with specific inputs, which we identify is not enough for generation tasks. To this end, in this paper, we propose a novel IP protection scheme for GANs where ownership verification can be done by checking outputs only, without choosing the inputs (i.e., box-free setting). Specifically, we make use of the unexploited potential of the discriminator to learn a hypersphere that captures the unique distribution learned by the paired generator. Extensive evaluations on two popular GAN tasks and more than 10 GAN architectures demonstrate our proposed scheme to effectively verify the ownership. Our proposed scheme shown to be immune to popular input-based removal attacks and robust against other existing attacks. The source code and models are available at https://github.com/AbstractTeen/gan_ownership_verification.
****
When Noisy Labels Meet Long Tail Dilemmas: A Representation Calibration Method
Manyi Zhang, Xuyang Zhao, Jun Yao, Chun Yuan, Weiran Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15890-15900
Real-world large-scale datasets are both noisily labeled and class-imbalanced. The issues seriously hurt the generalization of trained models. It is hence significant to address the simultaneous incorrect labeling and class-imbalance, i.e., the problem of learning with noisy labels on long-tailed data. Previous works develop several methods for the problem. However, they always rely on strong assumptions that are invalid or hard to be checked in practice. In this paper, to handle the problem and address the limitations of prior works, we propose a representation calibration method RCAL. Specifically, RCAL works with the representations extracted by unsupervised contrastive learning. We assume that without incorrect labeling and class imbalance, the representations of instances in each class conform to a multivariate Gaussian distribution, which is much milder and easier to be checked. Based on the assumption, we recover underlying representation distributions from polluted ones resulting from mislabeled and class-imbalanced data. Additional data points are then sampled from the recovered distributions to help generalization. Moreover, during classifier training, representation learning takes advantage of representation robustness brought by contrastive learning, which further improves the classifier performance. We derive theoretical results to discuss the effectiveness of our representation calibration. Experiments on multiple benchmarks justify our claims and confirm the superiority of the proposed method.
****
Reinforce Data, Multiply Impact: Improved Model Accuracy and Robustness with Dataset Reinforcement
Fartash Faghri, Hadi Pouransari, Sachin Mehta, Mehrdad Farajtabar, Ali Farhadi, Mohammad Rastegari, Oncel Tuzel; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17032-17043
We propose Dataset Reinforcement, a strategy to improve a dataset once such that the accuracy of any model architecture trained on the reinforced dataset is improved at no additional training cost for users. We propose a Dataset Reinforcement strategy based on data augmentation and knowledge distillation. Our generic strategy is designed based on extensive analysis across CNN- and transformer-based models and performing large-scale study of distillation with state-of-the-art models with various data augmentations. We create a reinforced version of the ImageNet training dataset, called ImageNet+, as well as reinforced datasets CIFAR-100+, Flowers-102+, and Food-101+. Models trained with ImageNet+ are more accurate, robust, and calibrated, and transfer well to downstream tasks (e.g., segmentation and detection). As an example, the accuracy of ResNet-50 improves by 1.7% on the ImageNet validation set, 3.5% on ImageNetV2, and 10.0% on ImageNet-R. Exp

ected Calibration Error (ECE) on the ImageNet validation set is also reduced by 9.9%. Using this backbone with Mask-RCNN for object detection on MS-COCO, the mean average precision improves by 0.8%. We reach similar gains for MobileNets, ViTs, and Swin-Transformers. For MobileNetV3 and Swin-Tiny, we observe significant improvements on ImageNet-R/A/C of up to 20% improved robustness. Models pretrained on ImageNet+ and fine-tuned on CIFAR-100+, Flowers-102+, and Food-101+, reach up to 3.4% improved accuracy. The code, datasets, and pretrained models are available at https://github.com/apple/ml-dr.
****
An Adaptive Model Ensemble Adversarial Attack for Boosting Adversarial Transferability
Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, Ximeng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4489-4498
While the transferability property of adversarial examples allows the adversary to perform black-box attacks i.e., the attacker has no knowledge about the target model), the transfer-based adversarial attacks have gained great attention. Previous works mostly study gradient variation or image transformations to amplify the distortion on critical parts of inputs. These methods can work on transferring across models with limited differences, i.e., from CNNs to CNNs, but always fail in transferring across models with wide differences, such as from CNNs to ViTs. Alternatively, model ensemble adversarial attacks are proposed to fuse outputs from surrogate models with diverse architectures to get an ensemble loss, making the generated adversarial example more likely to transfer to other models as it can fool multiple models concurrently. However, existing ensemble attacks simply fuse the outputs of the surrogate models evenly, thus are not efficacious to capture and amplify the intrinsic transfer information of adversarial examples. In this paper, we propose an adaptive ensemble attack, dubbed AdaEA, to adaptively control the fusion of the outputs from each model, via monitoring the discrepancy ratio of their contributions towards the adversarial objective. Furthermore, an extra disparity-reduced filter is introduced to further synchronize the update direction. As a result, we achieve considerable improvement over the existing ensemble attacks on various datasets, and the proposed AdaEA can also boost existing transfer-based attacks, which further demonstrates its efficacy and versatility.
****
Incremental Generalized Category Discovery
Bingchen Zhao, Oisin Mac Aodha; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19137-19147
We explore the problem of Incremental Generalized Category Discovery (IGCD). This is a challenging category-incremental learning setting where the goal is to develop models that can correctly categorize images from previously seen categories, in addition to discovering novel ones. Learning is performed over a series of time steps where the model obtains new labeled and unlabeled data, and discards old data, at each iteration. The difficulty of the problem is compounded in our generalized setting as the unlabeled data can contain images from categories that may or may not have been observed before. We present a new model for IGCD which combines non-parametric categorization with efficient image sampling to mitigate catastrophic forgetting. To quantify performance, we propose a new benchmark dataset named iNatIGCD that is motivated by a real-world fine-grained visual categorization task. In our experiments we outperform existing related methods.
****
Prototypical Mixing and Retrieval-Based Refinement for Label Noise-Resistant Image Retrieval
Xinlong Yang, Haixin Wang, Jinan Sun, Shikun Zhang, Chong Chen, Xian-Sheng Hua, Xiao Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11239-11249
Label noise is pervasive in real-world applications, which influences the optimization of neural network models. This paper investigates a realistic but understudied problem of image retrieval under label noise, which could lead to severe overfitting or memorization of noisy samples during optimization. Moreover, ident

ifying noisy samples correctly is still a challenging problem for retrieval models. In this paper, we propose a novel approach called Prototypical Mixing and Retrieval-based Refinement (TITAN) for label noise-resistant image retrieval, which corrects label noise and mitigates the effects of the memorization simultaneously. Specifically, we first characterize numerous prototypes with Gaussian distributions in the hidden space, which would direct the Mixing procedure in providing synthesized samples. These samples are fed into a similarity learning framework with varying emphasis based on the prototypical structure to learn semantics with reduced overfitting. In addition, we retrieve comparable samples for each prototype from simple to complex, which refine noisy samples in an accurate and class-balanced manner. Comprehensive experiments on five benchmark datasets demonstrate the superiority of our proposed TITAN compared with various competing baselines.
****

AccFlow: Backward Accumulation for Long-Range Optical Flow

Guangyang Wu, Xiaohong Liu, Kunming Luo, Xi Liu, Qingqing Zheng, Shuaicheng Liu, Xinyang Jiang, Guangtao Zhai, Wenyi Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12119-12128

Recent deep learning-based optical flow estimators have exhibited impressive performance in generating local flows between consecutive frames. However, the estimation of long-range flows between distant frames, particularly under complex object deformation and large motion occlusion, remains a challenging task. One promising solution is to accumulate local flows explicitly or implicitly to obtain the desired long-range flow. Nevertheless, the accumulation errors and flow misalignment can hinder the effectiveness of this approach. This paper proposes a novel recurrent framework called AccFlow, which recursively backward accumulates local flows using a deformable module called as AccPlus. In addition, an adaptive blending module is designed along with AccPlus to alleviate the occlusion effect by backward accumulation and rectify the accumulation error. Notably, we demonstrate the superiority of backward accumulation over conventional forward accumulation, which to the best of our knowledge has not been explicitly established before. To train and evaluate the proposed AccFlow, we have constructed a large-scale high-quality dataset named CVO, which provides ground-truth optical flow labels between adjacent and distant frames. Extensive experiments validate the effectiveness of AccFlow in handling long-range optical flow estimation. Codes are available at https://github.com/mulns/AccFlow.
****

Guiding Local Feature Matching with Surface Curvature

Shuzhe Wang, Juho Kannala, Marc Pollefeys, Daniel Barath; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17981-17991

We propose a new method, named curvature similarity extractor (CSE), for improving local feature matching across images. CSE calculates the curvature of the local 3D surface patch for each detected feature point in a viewpoint-invariant manner via fitting quadrics to predicted monocular depth maps. This curvature is then leveraged as an additional signal in feature matching with off-the-shelf matchers like SuperGlue and LoFTR. Additionally, CSE enables end-to-end joint training by connecting the matcher and depth predictor networks. Our experiments demonstrate on large-scale real-world datasets that CSE continuously improves the accuracy of state-of-the-art methods. Fine-tuning the depth prediction network further enhances the accuracy. The proposed approach achieves state-of-the-art results on the ScanNet dataset, showcasing the effectiveness of incorporating 3D geometric information into feature matching.
****

3D-VisTA: Pre-trained Transformer for 3D Vision and Text Alignment

Ziyu Zhu, Xiaojian Ma, Yixin Chen, Zhidong Deng, Siyuan Huang, Qing Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2911-2921

3D vision-language grounding (3D-VL) is an emerging field that aims to connect the 3D physical world with natural language, which is crucial for achieving embodied intelligence. Current 3D-VL models rely heavily on sophisticated modules, au

xiliary losses, and optimization tricks, which calls for a simple and unified model. In this paper, we propose 3D-VisTA, a pre-trained Transformer for 3D Vision and Text Alignment that can be easily adapted to various downstream tasks. 3D-VisTA simply utilizes self-attention layers for both single-modal modeling and multi-modal fusion without any sophisticated task-specific design. To further enhance its performance on 3D-VL tasks, we construct ScanScribe, the first large-scale 3D scene-text pairs dataset for 3D-VL pre-training. ScanScribe contains 2,995 RGB-D scans for 1,185 unique indoor scenes originating from ScanNet and 3R-Scan datasets, along with paired 278K scene descriptions generated from existing 3D-VL tasks, templates, and GPT-3. 3D-VisTA is pre-trained on ScanScribe via masked language/object modeling and scene-text matching. It achieves state-of-the-art results on various 3D-VL tasks, ranging from visual grounding and question answering to situated reasoning. Moreover, 3D-VisTA demonstrates superior data efficiency, obtaining strong performance even with limited annotations during downstream task fine-tuning.
****
Constraining Depth Map Geometry for Multi-View Stereo: A Dual-Depth Approach with Saddle-shaped Depth Cells
Xinyi Ye, Weiyue Zhao, Tianqi Liu, Zihao Huang, Zhiguo Cao, Xin Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17661-17670
Learning-based multi-view stereo (MVS) methods deal with predicting accurate depth maps to achieve an accurate and complete 3D representation. Despite the excellent performance, existing methods ignore the fact that a suitable depth geometry is also critical in MVS. In this paper, we demonstrate that different depth geometries have significant performance gaps, even using the same depth prediction error. Therefore, we introduce an ideal depth geometry composed of Saddle-Shaped Cells, whose predicted depth map oscillates upward and downward around the ground-truth surface, rather than maintaining a continuous and smooth depth plane. To achieve it, we develop a coarse-to-fine framework called Dual-MVSNet (DMVSNet), which can produce an oscillating depth plane. Technically, we predict two depth values for each pixel (Dual-Depth) and propose a novel loss function and a checkerboard-shaped selecting strategy to constrain the predicted depth geometry. Compared to existing methods, DMVSNet achieves a high rank on the DTU benchmark and obtains the top performance on challenging scenes of Tanks and Temples, demonstrating its strong performance and generalization ability. Our method also points to a new research direction for considering depth geometry in MVS.
****
SparseDet: Improving Sparsely Annotated Object Detection with Pseudo-positive Mining
Saksham Suri, Saketh Rambhatla, Rama Chellappa, Abhinav Shrivastava; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6770-6781
Training with sparse annotations is known to reduce the performance of object detectors. Previous methods have focused on proxies for missing ground truth annotations in the form of pseudo-labels for unlabeled boxes. We observe that existing methods suffer at higher levels of sparsity in the data due to noisy pseudo-labels. To prevent this, we propose an end-to-end system that learns to separate the proposals into labeled and unlabeled regions using Pseudo-positive mining. While the labeled regions are processed as usual, self-supervised learning is used to process the unlabeled regions thereby preventing the negative effects of noisy pseudo-labels. This novel approach has multiple advantages such as improved robustness to higher sparsity when compared to existing methods. We conduct exhaustive experiments on five splits on the PASCAL-VOC and COCO datasets achieving state-of-the-art performance. We also unify various splits used across literature for this task and present a standardized benchmark. On average, we improve by 2.6, 3.9 and 9.6 mAP over previous state-of-the-art methods on three splits of increasing sparsity on COCO. Our project is publicly available at cs.umd.edu/ sakshams/SparseDet.
****

Among Us: Adversarially Robust Collaborative Perception by Consensus
Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, Chen Feng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 186-195
Multiple robots could perceive a scene (e.g., detect objects) collaboratively better than individuals, although easily suffer from adversarial attacks when using deep learning. This could be addressed by the adversarial defense, but its training requires the often-unknown attacking mechanism. Differently, we propose ROBOSAC, a novel sampling-based defense strategy generalizable to unseen attackers. Our key idea is that collaborative perception should lead to consensus rather than dissensus in results compared to individual perception. This leads to our hypothesize-and-verify framework: perception results with and without collaboration from a random subset of teammates are compared until reaching a consensus. In such a framework, more teammates in the sampled subset often entail better perception performance but require longer sampling time to reject potential attackers. Thus, we derive how many sampling trials are needed to ensure the desired size of an attacker-free subset, or equivalently, the maximum size of such a subset that we can successfully sample within a given number of trials. We validate our method on the task of collaborative 3D object detection in autonomous driving scenarios.
****
BUS: Efficient and Effective Vision-Language Pre-Training with Bottom-Up Patch Summarization.
Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, Songfang Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2900-2910
Vision Transformer (ViT) based Vision-Language Pretraining (VLP) models recently demonstrated impressive performance in various tasks. However, the lengthy visual token sequences used in these models can lead to inefficient and ineffective performance. Existing methods to address these issues lack textual guidance and may overlook crucial visual information related to the text, leading to the introduction of irrelevant information during cross-modal fusion and additional computational cost. In this paper, we propose a Bottom-Up Patch Summarization approach named BUS which is inspired by the Document Summarization Task in NLP to learn a concise visual summary of lengthy visual token sequences, guided by textual semantics. We introduce a Text-Semantic Aware Patch Selector (TAPS) in the ViT backbone to perform a coarse-grained selective visual summarization to over-determine the text-relevant patches, and a light Summarization Decoder to perform fine-grained abstractive summarization based on the selected patches, resulting in a further condensed representation sequence that highlights text-relevant visual semantic information. Such bottom-up process is both efficient and effective with higher performing. We evaluate our approach on various VL understanding and generation tasks and show competitive or better downstream task performance while boosting the efficiency by 50%. Additionally, our model achieves well-designed SOTA downstream task performance by increasing input image resolution without increasing computational costs compared to baselines.
****
DiffusionDet: Diffusion Model for Object Detection
Shoufa Chen, Peize Sun, Yibing Song, Ping Luo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19830-19843
We propose DiffusionDet, a new framework that formulates object detection as a denoising diffusion process from noisy boxes to object boxes. During the training stage, object boxes diffuse from ground-truth boxes to random distribution, and the model learns to reverse this noising process. In inference, the model refines a set of randomly generated boxes to the output results in a progressive way. Our work possesses an appealing property of flexibility, which enables the dynamic number of boxes and iterative evaluation. The extensive experiments on the standard benchmarks show that DiffusionDet achieves favorable performance compared to previous well-established detectors. For example, DiffusionDet achieves 5.3 AP and 4.8 AP gains when evaluated with more boxes and iteration steps, under a

zero-shot transfer setting from COCO to CrowdHuman. Our code is available at https://github.com/ShoufaChen/DiffusionDet.
****

Forward Flow for Novel View Synthesis of Dynamic Scenes

Xiang Guo, Jiadai Sun, Yuchao Dai, Guanying Chen, Xiaoqing Ye, Xiao Tan, Errui Ding, Yumeng Zhang, Jingdong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16022-16033

This paper proposes a neural radiance field (NeRF) approach for novel view synthesis of dynamic scenes using forward warping. Existing methods often adopt a static NeRF to represent the canonical space, and render dynamic images at other time steps by mapping the sampled 3D points back to the canonical space with the learned backward flow field. However, this backward flow field is non-smooth and discontinuous, which is difficult to be fitted by commonly used smooth motion models. To address this problem, we propose to estimate the forward flow field and directly warp the canonical radiance field to other time steps. Such forward flow field is smooth and continuous within the object region, which benefits the motion model learning. To achieve this goal, we represent the canonical radiance field with voxel grids to enable efficient forward warping, and propose a differentiable warping process, including an average splatting operation and an inpaint network, to resolve the many-to-one and one-to-many mapping issues. Thorough experiments show that our method outperforms existing methods in both novel view rendering and motion modeling, demonstrating the effectiveness of our forward flow motion modeling. Project page: https://npucvr.github.io/ForwardFlowDNeRF.
****

CopyRNeRF: Protecting the CopyRight of Neural Radiance Fields

Ziyuan Luo, Qing Guo, Ka Chun Cheung, Simon See, Renjie Wan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22401-22411

Neural Radiance Fields (NeRF) have the potential to be a major representation of media. Since training a NeRF has never been an easy task, the protection of its model copyright should be a priority. In this paper, by analyzing the pros and cons of possible copyright protection solutions, we propose to protect the copyright of NeRF models by replacing the original color representation in NeRF with a watermarked color representation. Then, a distortion-resistant rendering scheme is designed to guarantee robust message extraction in 2D renderings of NeRF. Our proposed method can directly protect the copyright of NeRF models while maintaining high rendering quality and bit accuracy when compared among optional solutions.
****

Contrastive Model Adaptation for Cross-Condition Robustness in Semantic Segmentation

David Brüggemann, Christos Sakaridis, Tim Broedermann, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11378-11387

Standard unsupervised domain adaptation methods adapt models from a source to a target domain using labeled source data and unlabeled target data jointly. In model adaptation, on the other hand, access to the labeled source data is prohibited, i.e., only the source-trained model and unlabeled target data are available. We investigate normal-to-adverse condition model adaptation for semantic segmentation, whereby image-level correspondences are available in the target domain. The target set consists of unlabeled pairs of adverse- and normal-condition street images taken at GPS-matched locations. Our method--CMA--leverages such image pairs to learn condition-invariant features via contrastive learning. In particular, CMA encourages features in the embedding space to be grouped according to their condition-invariant semantic content and not according to the condition under which respective inputs are captured. To obtain accurate cross-domain semantic correspondences, we warp the normal image to the viewpoint of the adverse image and leverage warp-confidence scores to create robust, aggregated features. With this approach, we achieve state-of-the-art semantic segmentation performance for model adaptation on several normal-to-adverse adaptation benchmarks, such as

ACDC and Dark Zurich. We also evaluate CMA on a newly procured adverse-condition generalization benchmark and report favorable results compared to standard unsupervised domain adaptation methods, despite the comparative handicap of CMA due to source data inaccessibility. Code is available at https://github.com/brdav/cma.

****

SegRCDB: Semantic Segmentation via Formula-Driven Supervised Learning

Risa Shinoda, Ryo Hayamizu, Kodai Nakashima, Nakamasa Inoue, Rio Yokota, Hirokatsu Kataoka; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20054-20063

Pre-training is a strong strategy for enhancing visual models to efficiently train them with a limited number of labeled images. In semantic segmentation, creating annotation masks requires an intensive amount of labor and time, and therefore, a large-scale pre-training dataset with semantic labels is quite difficult to construct. Moreover, what matters in semantic segmentation pre-training has not been fully investigated. In this paper, we propose the Segmentation Radial Contour DataBase (SegRCDB), which for the first time applies formula-driven supervised learning for semantic segmentation. SegRCDB enables pre-training for semantic segmentation without real images or any manual semantic labels. SegRCDB is based on insights about what is important in pre-training for semantic segmentation and allows efficient pre-training. Pre-training with SegRCDB achieved higher mIoU than the pre-training with COCO-Stuff for fine-tuning on ADE-20k and Cityscapes with the same number of training images. SegRCDB has a high potential to contribute to semantic segmentation pre-training and investigation by enabling the creation of large datasets without manual annotation. The SegRCDB dataset will be released under a license that allows research and commercial use.

****

Creative Birds: Self-Supervised Single-View 3D Style Transfer

Renke Wang, Guimin Que, Shuo Chen, Xiang Li, Jun Li, Jian Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8775-8784

In this paper, we propose a novel method for single-view 3D style transfer that generates a unique 3D object with both shape and texture transfer.
 Our focus lies primarily on birds, a popular subject in 3D reconstruction, for which no existing single-view 3D transfer methods have been developed. The method we propose seeks to generate a 3D mesh shape and texture of a bird from two single-view images. To achieve this, we introduce a novel shape transfer generator that comprises a dual residual gated network (DRGNet), and a multi-layer perceptron (MLP). DRGNet extracts the features of source and target images using a shared coordinate gate unit, while the MLP generates spatial coordinates for building a 3D mesh. We also introduce a semantic UV texture transfer module that implements textural style transfer using semantic UV segmentation, which ensures consistency in the semantic meaning of the transferred regions. This module can be widely adapted to many existing approaches. Finally, our method constructs a novel 3D bird using a differentiable renderer. Experimental results on the CUB dataset verify that our method achieves state-of-the-art performance on the single-view 3D style transfer task. Code is available at https://github.com/wrk226/creative_birds.

****

LoTE-Animal: A Long Time-span Dataset for Endangered Animal Behavior Understanding

Dan Liu, Jin Hou, Shaoli Huang, Jing Liu, Yuxin He, Bochuan Zheng, Jifeng Ning, Jingdong Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20064-20075

Understanding and analyzing animal behavior is increasingly essential to protect endangered animal species. However, the application of advanced computer vision techniques in this regard is minimal, which boils down to lacking large and diverse datasets for training deep models. To break the deadlock, we present LoTE-Animal, a large-scale endangered animal dataset collected over 12 years, to foster the application of deep learning in rare species conservation. The collected d

ata contains vast variations such as ecological seasons, weather conditions, per iods, viewpoints, and habitat scenes. So far, we retrieved at least 500K videos and 1.2 million images. Specifically, we selected and annotated 11 endangered an imals for behavior understanding, including 10K video sequences for the action r ecognition task, 28K images for object detection, instance segmentation, and pos e estimation tasks. In addition, we gathered 7K web images of the same species a s source domain data for the domain adaptation task. We provide evaluation resul ts of representative vision understanding approaches and cross-domain experiment s. LoTE-Animal dataset would facilitate the community to research more advanced machine learning models and explore new tasks to aid endangered animal conservat ion. Our dataset will be released with the paper.
****
DQS3D: Densely-matched Quantization-aware Semi-supervised 3D Detection
Huan-ang Gao, Beiwen Tian, Pengfei Li, Hao Zhao, Guyue Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21905-219 15
In this paper, we study the problem of semi-supervised 3D object detection, whic h is of great importance considering the high annotation cost for cluttered 3D i ndoor scenes. We resort to the robust and principled framework of self-teaching, which has triggered notable progress for semi-supervised learning recently. Whi le this paradigm is natural for image-level or pixel-level prediction, adapting it to the detection problem is challenged by the issue of proposal matching. Pri or methods are based upon two-stage pipelines, matching heuristically selected p roposals generated in the first stage and resulting in spatially sparse training signals. In contrast, we propose the first semi-supervised 3D detection algorit hm that works in the single-stage manner and allows spatially dense training sig nals. A fundamental issue of this new design is the quantization error caused by point-to-voxel discretization, which inevitably leads to misalignment between t wo transformed views in the voxel domain. To this end, we derive and implement c losed-form rules that compensate this misalignment on-the-fly. Our results are s ignificant, e.g., promoting ScanNet mAP@0.5 from 35.2% to 48.5% using 20% annota tion. Codes and data are publicly available.
****
Towards Inadequately Pre-trained Models in Transfer Learning
Andong Deng, Xingjian Li, Di Hu, Tianyang Wang, Haoyi Xiong, Cheng-Zhong Xu; Pro ceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 202 3, pp. 19397-19408
Transfer learning has been a popular learning paradigm in the deep learning era, especially in annotation-insufficient scenarios. Better ImageNet pre-trained mo dels have been demonstrated, from the perspective of architecture, by previous r esearch to have better transferability to downstream tasks. However, in this pap er, we find that during the same pre-training process, models at middle epochs, which are inadequately pre-trained, can outperform fully trained models when use d as feature extractors (FE), while the fine-tuning (FT) performance still grows with the source performance. This reveals that there is not a solid positive co rrelation between top-1 accuracy on ImageNet and the transferring result on targ et data. Based on the contradictory phenomenon between FE and FT that a better f eature extractor fails to be fine-tuned better accordingly, we conduct comprehen sive analyses on features before the softmax layer to provide insightful explana tions. Our discoveries suggest that, during pre-training, models tend to first l earn spectral components corresponding to large singular values and the residual components contribute more when fine-tuning.
****
Boosting Novel Category Discovery Over Domains with Soft Contrastive Learning an d All in One Classifier
Zelin Zang, Lei Shang, Senqiao Yang, Fei Wang, Baigui Sun, Xuansong Xie, Stan Z. Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC CV), 2023, pp. 11858-11867
Unsupervised domain adaptation (UDA) has proven to be highly effective in transf erring knowledge from a label-rich source domain to a label-scarce target domain

. However, the presence of additional novel categories in the target domain has led to the development of open-set domain adaptation (ODA) and universal domain adaptation (UNDA). Existing ODA and UNDA methods treat all novel categories as a single, unified unknown class and attempt to detect it during training. However, we found that domain variance can lead to more significant view-noise in unsupervised data augmentation, which affects the effectiveness of contrastive learning (CL) and causes the model to be overconfident in novel category discovery. To address these issues, a framework nameded Soft-contrastive All-in-one Network (SAN) is proposed for ODA and UNDA tasks. SAN includes a novel data-augmentation-based soft contrastive learning (SCL) loss to fine-tune the backbone for feature transfer and a more human-intuitive classifier to improve new class discovery capability. The SCL loss weakens the adverse effects of the data augmentation view-noise problem which is amplified in domain transfer tasks. The All-in-One (AIO) classifier overcomes the overconfidence problem of current mainstream closed-set and open-set classifiers. Visualization and ablation experiments demonstrate the effectiveness of the proposed innovations. Furthermore, extensive experiment results on ODA and UNDA show that SAN outperforms existing state-of-the-art methods.
****
Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification
Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, Jun Cheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18905-18915
"A picture is worth a thousand words", significantly beyond mere a categorization. Accompanied by that, many patches of the image could have completely irrelevant meanings with the categorization if they were independently observed. This could significantly reduce the efficiency of a large family of few-shot learning algorithms, which have limited data and highly rely on the comparison of image patches. To address this issue, we propose a Class-aware Patch Embedding Adaptation (CPEA) method to learn "class-aware embeddings" of the image patches. The key idea of CPEA is to integrate patch embeddings with class-aware embeddings to make them class-relevant. Furthermore, we define a dense score matrix between class-relevant patch embeddings across images, based on which the degree of similarity between paired images is quantified. Visualization results show that CPEA concentrates patch embeddings by class, thus making them class-relevant. Extensive experiments on four benchmark datasets, miniImageNet, tieredImageNet, CIFAR-FS, and FC-100, indicate that our CPEA significantly outperforms the existing state-of-the-art methods. The source code is available at https://github.com/FushengHao/CPEA.
****
SegPrompt: Boosting Open-World Segmentation via Category-Level Prompt Learning
Muzhi Zhu, Hengtao Li, Hao Chen, Chengxiang Fan, Weian Mao, Chenchen Jing, Yifan Liu, Chunhua Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 999-1008
Current closed-set instance segmentation models rely on predefined class labels for each mask during training and evaluation, limiting their ability to detect novel objects. Open-world instance segmentation (OWIS) models address this challenge by detecting unknown objects in a class-agnostic manner. However, previous OWIS approaches completely erase category information during training to keep the model's ability to generalize to unknown objects. In this work, we propose a novel training mechanism called SegPrompt that utilizes category information to improve the model's class-agnostic segmentation ability for both known and unknown categories. In addition, the previous OWIS training setting exposes the unknown classes to the training set and brings information leakage, which is unreasonable in the real world. Therefore, we provide a new open-world benchmark closer to a real-world scenario by dividing the dataset classes into known-seen-unseen parts. For the first time we focus on the model's ability to discover objects that never appear in the training set images. Experiments show that SegPrompt can improve the overall and unseen detection performance by 5.6% and 6.1% in AR on our new benchmark without affecting the inference efficiency. We further demonstrat

e the effectiveness of our method on existing cross-dataset transfer and strongl y supervised settings, leading to 5.5% and 12.3% relative improvement.
****
Search for or Navigate to? Dual Adaptive Thinking for Object Navigation
Ronghao Dang, Liuyi Wang, Zongtao He, Shuai Su, Jiagui Tang, Chengju Liu, Qijun Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8250-8259
"Search for" or "Navigate to"? When we find a specific object in an unknown environment, the two choices always arise in our subconscious mind. Before we see the target, we search for the target based on prior experience. Once we have seen the target, we can navigate to it by remembering the target location. However, recent object navigation methods consider using object association mostly to enhance the "search for" phase while neglecting the importance of the "navigate to" phase. Therefore, this paper proposes a dual adaptive thinking (DAT) method that flexibly adjusts thinking strategies in different navigation stages. Dual thinking includes both search thinking according to the object association ability and navigation thinking according to the target location ability. To make navigation thinking more effective, we design a target-oriented memory graph (TOMG) (which stores historical target information) and a target-aware multi-scale aggregator (TAMSA) (which encodes the relative position of the target). We assess our methods based on the AI2-Thor and RoboTHOR datasets. Compared with state-of-the-art (SOTA) methods, our approach significantly raises the overall success rate (SR) and success weighted by path length (SPL) while enhancing the agent's performance in the "navigate to" phase.
****
CL-MVSNet: Unsupervised Multi-View Stereo with Dual-Level Contrastive Learning
Kaiqiang Xiong, Rui Peng, Zhe Zhang, Tianxing Feng, Jianbo Jiao, Feng Gao, Ronggang Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3769-3780
Unsupervised Multi-View Stereo (MVS) methods have achieved promising progress recently. However, previous methods primarily depend on the photometric consistency assumption, which may suffer from two limitations: indistinguishable regions and view-dependent effects, e.g., low-textured areas and reflections. To address these issues, in this paper, we propose a new dual-level contrastive learning approach, named CL-MVSNet. Specifically, our model integrates two contrastive branches into an unsupervised MVS framework to construct additional supervisory signals. On the one hand, we present an image-level contrastive branch to guide the model to acquire more context awareness, thus leading to more complete depth estimation in indistinguishable regions. On the other hand, we exploit a scene-level contrastive branch to boost the representation ability, improving robustness to view-dependent effects. Moreover, to recover more accurate 3D geometry, we introduce an L0.5 photometric consistency loss, which encourages the model to focus more on accurate points while mitigating the gradient penalty of undesirable ones. Extensive experiments on DTU and Tanks&Temples benchmarks demonstrate that our approach achieves state-of-the-art performance among all end-to-end unsupervised MVS frameworks and outperforms its supervised counterpart by a considerable margin without fine-tuning.
****
Federated Learning Over Images: Vertical Decompositions and Pre-Trained Backbones Are Difficult to Beat
Erdong Hu, Yuxin Tang, Anastasios Kyrillidis, Chris Jermaine; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19385-19396
We carefully evaluate a number of algorithms for learning in a federated environment, and test their utility for a variety of image classification tasks. We consider many issues that have not been adequately considered before: whether learning over data sets that do not have diverse sets of images affects the results; whether to use a pre-trained feature extraction "backbone"; how to evaluate learner performance (we argue that classification accuracy is not enough), among others. Overall, across a wide variety of settings, we find that vertically decompo

sing a neural network seems to give the best results, and outperforms more stand ard reconciliation-used methods.
****
HOSNeRF: Dynamic Human-Object-Scene Neural Radiance Fields from a Single Video
Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, Mike Zheng Shou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18483-18494
We introduce HOSNeRF, a novel 360deg free-viewpoint rendering method that reconstructs neural radiance fields for dynamic human-object-scene from a single monocular in-the-wild video. Our method enables pausing the video at any frame and rendering all scene details (dynamic humans, objects, and backgrounds) from arbitrary viewpoints. The first challenge in this task is the complex object motions in human-object interactions, which we tackle by introducing the new object bones into the conventional human skeleton hierarchy to effectively estimate large object deformations in our dynamic human-object model. The second challenge is that humans interact with different objects at different times, for which we introduce two new learnable object state embeddings that can be used as conditions for learning our human-object representation and scene representation, respectively. Extensive experiments show that HOSNeRF significantly outperforms SOTA approaches on two challenging datasets by a large margin of 40%   50% in terms of LPIPS. The code, data, and compelling examples of 360deg free-viewpoint renderings from single videos: https://showlab.github.io/HOSNeRF.
****
OmniZoomer: Learning to Move and Zoom in on Sphere at High-Resolution
Zidong Cao, Hao Ai, Yan-Pei Cao, Ying Shan, Xiaohu Qie, Lin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12897-12907
Omnidirectional images (ODIs) have become increasingly popular, as their large field-of-view (FoV) can offer viewers the chance to freely choose the view directions in immersive environments such as virtual reality. The Mobius transformation is typically employed to further provide the opportunity for movement and zoom on ODIs, but applying it to the image level often results in blurry effect and aliasing problem. In this paper, we propose a novel deep learning-based approach, called OmniZoomer, to incorporate the Mobius transformation into the network for movement and zoom on ODIs. By learning various transformed feature maps under different conditions, the network is enhanced to handle the increasing edge curvatures, which alleviates the blurry effect. Moreover, to address the aliasing problem, we propose two key components. Firstly, to compensate for the lack of pixels for describing curves, we enhance the feature maps in the high-resolution (HR) space and calculate the transformed index map with a spatial index generation module. Secondly, considering that ODIs are inherently represented in the spherical space, we propose a spherical resampling module that combines the index map and HR feature maps to transform the feature maps for better spherical correlation. The transformed feature maps are decoded to output a zoomed ODI. Experiments show that our method can produce HR and high-quality ODIs with the flexibility to move and zoom in to the object of interest. Project page is available at http://vlislab22.github.io/OmniZoomer/.
****
Knowing Where to Focus: Event-aware Transformer for Video Grounding
Jinhyun Jang, Jungin Park, Jin Kim, Hyeongjun Kwon, Kwanghoon Sohn; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13846-13856
Recent DETR-based video grounding models have made the model directly predict moment timestamps without any hand-crafted components, such as a pre-defined proposal or non-maximum suppression, by learning moment queries.
However, their input-agnostic moment queries inevitably overlook an intrinsic temporal structure of a video, providing limited positional information.
In this paper, we formulate an event-aware dynamic moment query to enable the model to take the input-specific content and positional information of the video into account. To this end, we present two levels of reasoning: 1) Event reasonin

g that captures distinctive event units constituting a given video using a slot attention mechanism; and 2) moment reasoning that fuses the moment queries with a given sentence through a gated fusion transformer layer and learns interaction s between the moment queries and video-sentence representations to predict momen t timestamps.
 Extensive experiments demonstrate the effectiveness and efficiency of the event -aware dynamic moment queries, outperforming state-of-the-art approaches on seve ral video grounding benchmarks. The code is publicly available at https://github .com/jinhyunj/EaTR.
****
TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition
Shilin Lu, Yanzhu Liu, Adams Wai-Kin Kong; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2023, pp. 2294-2305
Text-driven diffusion models have exhibited impressive generative capabilities, enabling various image editing tasks. In this paper, we propose TF-ICON, a novel Training-Free Image COmpositioN framework that harnesses the power of text-driv en diffusion models for cross-domain image-guided composition. This task aims to seamlessly integrate user-provided objects into a specific visual context. Curr ent diffusion-based methods often involve costly instance-based optimization or finetuning of pretrained models on customized datasets, which can potentially un dermine their rich prior. In contrast, TF-ICON can leverage off-the-shelf diffus ion models to perform cross-domain image-guided composition without requiring ad ditional training, finetuning, or optimization. Moreover, we introduce the excep tional prompt, which contains no information, to facilitate text-driven diffusio n models in accurately inverting real images into latent representations, formin g the basis for compositing. Our experiments show that equipping Stable Diffusio n with the exceptional prompt outperforms state-of-the-art inversion methods on various datasets (CelebA-HQ, COCO, and ImageNet), and that TF-ICON surpasses pri or baselines in versatile visual domains. Code is available at https://github.co m/Shilin-LU/TF-ICON
****
Landscape Learning for Neural Network Inversion
Ruoshi Liu, Chengzhi Mao, Purva Tendulkar, Hao Wang, Carl Vondrick; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22 39-2250
Many machine learning methods operate by inverting a neural network at inference time, which has become a popular technique for solving inverse problems in comp uter vision, robotics, and graphics. However, these methods often involve gradie nt descent through a highly non-convex loss landscape, causing the optimization process to be unstable and slow. We introduce a method that learns a loss landsc ape where gradient descent is efficient, bringing massive improvement and accele ration to the inversion process. We demonstrate this advantage on a number of me thods for both generative and discriminative tasks, including GAN inversion, adv ersarial defense, and 3D human pose reconstruction.
****
Movement Enhancement toward Multi-Scale Video Feature Representation for Tempora l Action Detection
Zixuan Zhao, Dongqi Wang, Xu Zhao; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2023, pp. 13555-13564
Boundary localization is a challenging problem in Temporal Action Detection (TAD), in which there are two main issues. First, the submergence of movement featur e, i.e. the movement information in a snippet is covered by the scene informatio n. Second, the scale of action, that is, the proportion of action segments in th e entire video, is considerably variable. In this work, we first design a Moveme nt Enhance Module (MEM) to highlight movement feature for better action location , and then, we propose a Scale Feature Pyramid Network (SFPN) to detect multi-sc ale actions in videos. For Movement Enhance Module, firstly, Movement Feature Ex tractor (MFE) is designed to get the movement feature. Secondly, we propose a Mu lti-Relation Enhance Module (MREM) to grasp valuable information correlation bot h locally and temporally. For Scale Feature Pyramid Network, we design a U-Shape

Module to model different scale actions, moreover, we design the training and i nference strategy of different scales, ensuring that each pyramid layer is only responsible for actions at a specific scale. These two innovations are integrate d as the Movement Enhance Network (MENet), and extensive experiments conducted o n two challenging benchmarks demonstrate its effectiveness. MENet outperforms ot her representative TAD methods on ActivityNet-1.3 and THUMOS-14.
****
Collaborative Propagation on Multiple Instance Graphs for 3D Instance Segmentati on with Single-point Supervision
Shichao Dong, Ruibo Li, Jiacheng Wei, Fayao Liu, Guosheng Lin; Proceedings of th e IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16665-1 6674
Instance segmentation on 3D point clouds has been attracting increasing attentio n due to its wide applications, especially in scene understanding areas. However , most existing methods operate on fully annotated data while manually preparing ground-truth labels at point-level is very cumbersome and labor-intensive. To a ddress this issue, we propose a novel weakly supervised method RWSeg that only r equires labeling one object with one point. With these sparse weak labels, we in troduce a unified framework with two branches to propagate semantic and instance information respectively to unknown regions using self-attention and a cross-gr aph random walk method. Specifically, we propose a Cross-graph Competing Random Walks (CRW) algorithm that encourages competition among different instance graph s to resolve ambiguities in closely placed objects, improving instance assignmen t accuracy. RWSeg generates high-quality instance-level pseudo labels. Experimen tal results on ScanNet-v2 and S3DIS datasets show that our approach achieves com parable performance with fully-supervised methods and outperforms previous weakl y-supervised methods by a substantial margin.
****
PPR: Physically Plausible Reconstruction from Monocular Videos
Gengshan Yang, Shuo Yang, John Z. Zhang, Zachary Manchester, Deva Ramanan; Proce edings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3914-3924
Given monocular videos, we build 3D models of articulated objects and environmen ts whose 3D configurations satisfy dynamics and contact constraints. At its core , our method leverages differentiable physics simulation to aid visual reconstru ctions. We couple differentiable physics simulation with differentiable renderin g via coordinate descent, which enables end-to-end optimization of, not only 3D reconstructions, but also physical system parameters from videos. We demonstrate the effectiveness of physics-informed reconstruction on monocular videos of qua druped animals and humans. It reduces reconstruction artifacts (e.g., scale ambi guity, unbalanced poses, and foot swapping) that are challenging to address by v isual cues alone, and produces better foot contact estimation.
****
Single Image Deblurring with Row-dependent Blur Magnitude
Xiang Ji, Zhixiang Wang, Shin'ichi Satoh, Yinqiang Zheng; Proceedings of the IEE E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12269-12280
Image degradation often occurs during fast camera or object movements, regardles s of the exposure modes: global shutter (GS) or rolling shutter (RS). Since thes e two exposure modes give rise to intrinsically different degradations, two rest oration threads have been explored separately, i.e. motion deblurring of GS imag es and distortion correction of RS images, both of which are challenging restora tion tasks, especially in the presence of a single input image. In this paper, we explore a novel in-between exposure mode, called global reset release (GRR) shutter, which produces GS-like blur but with row-dependent blur magnitude. We t ake advantage of this unique characteristic of GRR to explore the latent frames within a single image and restore a clear counterpart by only relying on these latent contexts. Specifically, we propose a residual spatially-compensated and s pectrally-enhanced Transformer (RSS-T) block for row-dependent deblurring of a single GRR image. Its hierarchical positional encoding compensates global pos itional context of windows and enables order-awareness of the local pixel's posi

tion, along with a novel feed-forward network that simultaneously uses spatial a
nd spectral information for gaining mixed global context. Extensive experimental
 results demonstrate that our method outperforms the state-of-the-art GS deblurr
ing and RS correction methods on single GRR input.
****
Robust Heterogeneous Federated Learning under Data Corruption
Xiuwen Fang, Mang Ye, Xiyuan Yang; Proceedings of the IEEE/CVF International Con
ference on Computer Vision (ICCV), 2023, pp. 5020-5030
Model heterogeneous federated learning is a realistic and challenging problem. H
owever, due to the limitations of data collection, storage, and transmission con
ditions, as well as the existence of free-rider participants, the clients may su
ffer from data corruption. This paper starts the first attempt to investigate th
e problem of data corruption in the model heterogeneous federated learning frame
work. We design a novel method named Augmented Heterogeneous Federated Learning
(AugHFL), which consists of two stages: 1) In the local update stage, a corrupti
on-robust data augmentation strategy is adopted to minimize the adverse effects
of local corruption while enabling the models to learn rich local knowledge. 2)
In the collaborative update stage, we design a robust re-weighted communication
approach, which implements communication between heterogeneous models while miti
gating corrupted knowledge transfer from others. Extensive experiments demonstra
te the effectiveness of our method in coping with various corruption patterns in
 the model heterogeneous federated learning setting.
****
RMP-Loss: Regularizing Membrane Potential Distribution for Spiking Neural Networ
ks
Yufei Guo, Xiaode Liu, Yuanpei Chen, Liwen Zhang, Weihang Peng, Yuhan Zhang, Xuh
ui Huang, Zhe Ma; Proceedings of the IEEE/CVF International Conference on Comput
er Vision (ICCV), 2023, pp. 17391-17401
Spiking Neural Networks (SNNs) as one of the biology-inspired models have receiv
ed much attention recently. It can significantly reduce energy consumption since
 they quantize the real-valued membrane potentials to 0/1 spikes to transmit inf
ormation thus the multiplications of activations and weights can be replaced by
additions when implemented on hardware. However, this quantization mechanism wil
l inevitably introduce quantization error, thus causing catastrophic information
 loss. To address the quantization error problem, we propose a regularizing memb
rane potential loss (RMP-Loss) to adjust the distribution which is directly rela
ted to quantization error to a range close to the spikes. Our method is extremel
y simple to implement and straightforward to train an SNN. Furthermore, it is sh
own to consistently outperform previous state-of-the-art methods over different
network architectures and datasets.
****
Cyclic-Bootstrap Labeling for Weakly Supervised Object Detection
Yufei Yin, Jiajun Deng, Wengang Zhou, Li Li, Houqiang Li; Proceedings of the IEE
E/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7008-7018
Recent progress in weakly supervised object detection is featured by a combinati
on of multiple instance detection networks (MIDN) and ordinal online refinement.
 However, with only image-level annotation, MIDN inevitably assigns high scores
to some unexpected region proposals when generating pseudo labels. These inaccur
ate high-scoring region proposals will mislead the training of subsequent refine
ment modules and thus hamper the detection performance. In this work, we explore
 how to ameliorate the quality of pseudo-labeling in MIDN. Formally, we devise C
yclic-Bootstrap Labeling (CBL), a novel weakly supervised object detection pipel
ine, which optimizes MIDN with rank information from a reliable teacher network.
 Specifically, we obtain this teacher network by introducing a weighted exponent
ial moving average strategy to take advantage of various refinement modules. A n
ovel class-specific ranking distillation algorithm is proposed to leverage the o
utput of weighted ensembled teacher network for distilling MIDN with rank inform
ation. As a result, MIDN is guided to assign higher scores to accurate proposals
, which further benefits final detection. Extensive experiments on the prevalent
 PASCAL VOC 2007 & 2012 and COCO datasets demonstrate the superior performance o

f our CBL framework.
****
Deep Active Contours for Real-time 6-DoF Object Tracking
Long Wang, Shen Yan, Jianan Zhen, Yu Liu, Maojun Zhang, Guofeng Zhang, Xiaowei Zhou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14034-14044
This paper solves the problem of real-time 6-DoF object tracking from an RGB video. Prior optimization-based methods optimize the object pose by aligning the projected model to the image based on handcrafted features, which are prone to suboptimal solutions. Recent learning-based methods use neural networks to predict the pose, which suffer from limited generalizability or computational efficiency. We propose a learning-based active contour model to make the best use of both worlds. Specifically, given an initial pose, we project the object model to the image plane to obtain the initial contour and use a lightweight network to predict how the contour should move to match the true object boundary, which provides the gradients to optimize the object pose. We also devise an efficient optimization algorithm to train our model end-to-end with pose supervision. Experimental results on semi-synthetic and real-world 6-DoF object tracking datasets demonstrate that our model outperforms state-of-the-art methods by a substantial margin in pose accuracy, while achieving real-time performance on mobile devices. Code is available on our project page: https://zju3dv.github.io/deep_ac/.
****
Tangent Sampson Error: Fast Approximate Two-view Reprojection Error for Central Camera Models
Mikhail Terekhov, Viktor Larsson; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3370-3378
In this paper we introduce the Tangent Sampson error, which is a generalization of the classical Sampson error in two-view geometry that allows for arbitrary central camera models. It only requires local gradients of the distortion map at the original correspondences (allowing for pre-computation) resulting in a negligible increase in computational cost when used in RANSAC or local refinement. The error effectively approximates the true-reprojection error for a large variety of cameras, including extremely wide field-of-view lenses that cannot be undistorted to a single pinhole image. We show experimentally that the new error outperforms competing approaches both when used for model scoring in RANSAC and for non-linear refinement of the relative camera pose.
****
Multi-grained Temporal Prototype Learning for Few-shot Video Object Segmentation
Nian Liu, Kepan Nan, Wangbo Zhao, Yuanwei Liu, Xiwen Yao, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Junwei Han, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18862-18871
Few-Shot Video Object Segmentation (FSVOS) aims to segment objects in a query video with the same category defined by a few annotated support images. However, this task was seldom explored. In this work, based on IPMT, a state-of-the-art few-shot image segmentation method that combines external support guidance information with adaptive query guidance cues, we propose to leverage multi-grained temporal guidance information for handling the temporal correlation nature of video data. We decompose the query video information into a clip prototype and a memory prototype for capturing local and long-term internal temporal guidance, respectively. Frame prototypes are further used for each frame independently to handle fine-grained adaptive guidance and enable bidirectional clip-frame prototype communication. To reduce the influence of noisy memory, we propose to leverage the structural similarity relation among different predicted regions and the support for selecting reliable memory frames. Furthermore, a new segmentation loss is also proposed to enhance the category discriminability of the learned prototypes. Experimental results demonstrate that our proposed video IPMT model significantly outperforms previous models on two benchmark datasets. Code is available at https://github.com/nankepan/VIPMT.
****

Improving 3D Imaging with Pre-Trained Perpendicular 2D Diffusion Models
Suhyeon Lee, Hyungjin Chung, Minyoung Park, Jonghyuk Park, Wi-Sun Ryu, Jong Chul Ye; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10710-10720
Diffusion models have become a popular approach for image generation and reconstruction due to their numerous advantages. However, most diffusion-based inverse problem-solving methods only deal with 2D images, and even recently published 3D methods do not fully exploit the 3D distribution prior. To address this, we propose a novel approach using two perpendicular pre-trained 2D diffusion models to solve the 3D inverse problem. By modeling the 3D data distribution as a product of 2D distributions sliced in different directions, our method effectively addresses the curse of dimensionality. Our experimental results demonstrate that our method is highly effective for 3D medical image reconstruction tasks, including MRI Z-axis super-resolution, compressed sensing MRI, and sparse-view CT. Our method can generate high-quality voxel volumes suitable for medical applications.
****
Time Does Tell: Self-Supervised Time-Tuning of Dense Image Representations
Mohammadreza Salehi, Efstratios Gavves, Cees G.M. Snoek, Yuki M. Asano; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16536-16547
Spatially dense self-supervised learning is a rapidly growing problem domain with promising applications for unsupervised segmentation and pretraining for dense downstream tasks. Despite the abundance of temporal data in the form of videos, this information-rich source has been largely overlooked. Our paper aims to address this gap by proposing a novel approach that incorporates temporal consistency in dense self-supervised learning. While methods designed solely for images face difficulties in achieving even the same performance on videos, our method improves not only the representation quality for videos - but also images. Our approach, which we call time-tuning, starts from image-pretrained models and fine-tunes them with a novel self-supervised temporal-alignment clustering loss on unlabeled videos. This effectively facilitates the transfer of high-level information from videos to image representations. Time-tuning improves the state-of-the-art by 8-10% for unsupervised semantic segmentation on videos and matches it for images. We believe this method paves the way for further self-supervised scaling by leveraging the abundant availability of videos. The implementation can be found here : https://github.com/SMSD75/Timetuning
****
CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow
Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, Jerome Revaud; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17969-17980
Despite impressive performance for high-level downstream tasks, self-supervised pre-training methods have not yet fully delivered on dense geometric vision tasks such as stereo matching or optical flow. The application of self-supervised concepts, such as instance discrimination or masked image modeling, to geometric tasks is an active area of research. In this work, we build on the recent cross-view completion framework, a variation of masked image modeling that leverages a second view from the same scene which makes it well suited for binocular downstream tasks. The applicability of this concept has so far been limited in at least two ways: (a) by the difficulty of collecting real-world image pairs -- in practice only synthetic data have been used -- and (b) by the lack of generalization of vanilla transformers to dense downstream tasks for which relative position is more meaningful than absolute position. We explore three avenues of improvement. First, we introduce a method to collect suitable real-world image pairs at large scale. Second, we experiment with relative positional embeddings and show that they enable vision transformers to perform substantially better. Third, we scale up vision transformer based cross-completion architectures, which is made possible by the use of large amounts of data. With these improvements, we show for

 the first time that state-of-the-art results on stereo matching and optical flo
w can be reached without using any classical task-specific techniques like corre
lation volume, iterative estimation, image warping or multi-scale reasoning, thu
s paving the way towards universal vision models.
****
ExBluRF: Efficient Radiance Fields for Extreme Motion Blurred Images
Dongwoo Lee, Jeongtaek Oh, Jaesung Rim, Sunghyun Cho, Kyoung Mu Lee; Proceedings
 of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1
7639-17648
We present ExBluRF, a novel view synthesis method for extreme motion blurred ima
ges based on efficient radiance fields optimization. Our approach consists of tw
o main components: 6-DOF camera trajectory-based motion blur formulation and vox
el-based radiance fields. From extremely blurred images, we optimize the sharp r
adiance fields by jointly estimating the camera trajectories that generate the b
lurry images. In training, multiple rays along the camera trajectory are accumul
ated to reconstruct single blurry color, which is equivalent to the physical mot
ion blur operation. We minimize the photo-consistency loss on blurred image spac
e and obtain the sharp radiance fields with camera trajectories that explain the
 blur of all images. The joint optimization on the blurred image space demands p
ainfully increasing computation and resources proportional to the blur size. Our
 method solves this problem by replacing the MLP-based framework to low-dimensio
nal 6-DOF camera poses and voxel-based radiance fields. Compared with the existi
ng works, our approach restores much sharper 3D scenes from challenging motion b
lurred views with the order of 10x less training time and GPU memory consumption
.
****
MPCViT: Searching for Accurate and Efficient MPC-Friendly Vision Transformer wit
h Heterogeneous Attention
Wenxuan Zeng, Meng Li, Wenjie Xiong, Tong Tong, Wen-jie Lu, Jin Tan, Runsheng Wa
ng, Ru Huang; Proceedings of the IEEE/CVF International Conference on Computer V
ision (ICCV), 2023, pp. 5052-5063
Secure multi-party computation (MPC) enables computation directly on encrypted d
ata and protects both data and model privacy in deep learning inference. However
, existing neural network architectures, including Vision Transformers (ViTs), a
re not designed or optimized for MPC and incur significant latency overhead. We
observe Softmax accounts for the major latency bottleneck due to a high communic
ation complexity, but can be selectively replaced or linearized without compromi
sing the model accuracy. Hence, in this paper, we propose an MPC-friendly ViT, d
ubbed MPCViT, to enable accurate yet efficient ViT inference in MPC. Based on a
systematic latency and accuracy evaluation of the Softmax attention and other at
tention variants, we propose a heterogeneous attention optimization space. We al
so develop a simple yet effective MPC-aware neural architecture search algorithm
 for fast Pareto optimization. To further boost the inference efficiency, we pro
pose MPCViT+, to jointly optimize the Softmax attention and other network compon
ents, including GeLU, matrix multiplication, etc. With extensive experiments, we
 demonstrate that MPCViT achieves 1.9%, 1.3% and 3.6% higher accuracy with 6.2x,
 2.9x and 1.9x latency reduction compared with baseline ViT, MPCFormer and THE-X
 on the Tiny-ImageNet dataset, respectively. MPCViT+ further achieves a better P
areto front compared with MPCViT. The code and models for evaluation are availab
le at https://github.com/PKU-SEC-Lab/mpcvit.
****
Online Class Incremental Learning on Stochastic Blurry Task Boundary via Mask an
d Visual Prompt Tuning
Jun-Yeong Moon, Keon-Hee Park, Jung Uk Kim, Gyeong-Moon Park; Proceedings of the
 IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11731-11
741
Continual learning aims to learn a model from a continuous stream of data, but i
t mainly assumes a fixed number of data and tasks with clear task boundaries. Ho
wever, in real-world scenarios, the number of input data and tasks is constantly
 changing in a statistical way, not a static way. Although recently introduced i

ncremental learning scenarios having blurry task boundaries somewhat address the
 above issues, they still do not fully reflect the statistical properties of rea
l-world situations because of the fixed ratio of disjoint and blurry samples. In
 this paper, we propose a new Stochastic incremental Blurry task boundary scenar
io, called Si-Blurry, which reflects the stochastic properties of the real-world
. We find that there are two major challenges in the Si-Blurry scenario: (1) int
er- and intra-task forgettings and (2) class imbalance problem. To alleviate the
m, we introduce Mask and Visual Prompt tuning (MVP). In MVP, to address the inte
r- and intra-task forgetting issues, we propose a novel instance-wise logit mask
ing and contrastive visual prompt tuning loss. Both of them help our model disce
rn the classes to be learned in the current batch. It results in consolidating t
he previous knowledge. In addition, to alleviate the class imbalance problem, we
 introduce a new gradient similarity-based focal loss and adaptive feature scali
ng to ease overfitting to the major classes and underfitting to the minor classe
s. Extensive experiments show that our proposed MVP significantly outperforms th
e existing state-of-the-art methods in our challenging Si-Blurry scenario.
****
Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators
Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhan
gyang Wang, Shant Navasardyan, Humphrey Shi; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2023, pp. 15954-15964
Recent text-to-video generation approaches rely on computationally heavy trainin
g and require large-scale video datasets. In this paper, we introduce a new task
, zero-shot text-to-video generation, and propose a low-cost approach (without a
ny training or optimization) by leveraging the power of existing text-to-image s
ynthesis methods (e.g., Stable Diffusion), making them suitable for the video do
main.
  Our key modifications include (i) enriching the latent codes of the generated
frames with motion dynamics to keep the global scene and the background time con
sistent; and (ii) reprogramming frame-level self-attention using a new cross-fra
me attention of each frame on the first frame, to preserve the context, appearan
ce, and identity of the foreground object.
  Experiments show that this leads to low overhead, yet high-quality and remarka
bly consistent video generation. Moreover, our approach is not limited to text-t
o-video synthesis but is also applicable to other tasks such as conditional and
content-specialized video generation, and Video Instruct-Pix2Pix, i.e., instruct
ion-guided video editing.
  As experiments show, our method performs comparably or sometimes better than r
ecent approaches, despite not being trained on additional video data. Our code i
s publicly available at: https://github.com/Picsart-AI-Research/Text2Video-Zero
.
****
Masked Spiking Transformer
Ziqing Wang, Yuetong Fang, Jiahang Cao, Qiang Zhang, Zhongrui Wang, Renjing Xu;
Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),
2023, pp. 1761-1771
The combination of Spiking Neural Networks (SNNs) and Transformers has attracted
 significant attention due to their potential for high energy efficiency and hig
h-performance nature. However, existing works on this topic typically rely on di
rect training, which can lead to suboptimal performance. To address this issue,
we propose to leverage the benefits of the ANN-to-SNN conversion method to combi
ne SNNs and Transformers, resulting in significantly improved performance over e
xisting state-of-the-art SNN models. Furthermore, inspired by the quantal synapt
ic failures observed in the nervous system, which reduce the number of spikes tr
ansmitted across synapses, we introduce a novel Masked Spiking Transformer (MST)
 framework. This incorporates a Random Spike Masking (RSM) method to prune redun
dant spikes and reduce energy consumption without sacrificing performance. Our e
xperimental results demonstrate that the proposed MST model achieves a significa
nt reduction of 26.8% in power consumption when the masking ratio is 75% while m
aintaining the same level of performance as the unmasked model. The code is avai

lable at: https://github.com/bic-L/Masked-Spiking-Transformer.
****
Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives
Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, Weisi Lin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20144-20154
The rapid increase in user-generated-content (UGC) videos calls for the development of effective video quality assessment (VQA) algorithms. However, the objective of the UGC-VQA problem is still ambiguous and can be viewed from two perspectives: the technical perspective, measuring the perception of distortions; and the aesthetic perspective, which relates to preference and recommendation on contents. To understand how these two perspectives affect overall subjective opinions in UGC-VQA, we conduct a large-scale subjective study to collect human quality opinions on overall quality of videos as well as perceptions from aesthetic and technical perspectives. The collected Disentangled Video Quality Database (DIVIDE-3k) confirms that human quality opinions on UGC videos are universally and inevitably affected by both aesthetic and technical perspectives. In light of this, we propose the Disentangled Objective Video Quality Evaluator (DOVER) to learn the quality of UGC videos based on the two perspectives. The DOVER proves state-of-the-art performance in UGC-VQA under very high efficiency. With perspective opinions in DIVIDE-3k, we further propose DOVER++, the first approach to provide reliable clear-cut quality evaluations from a single aesthetic or technical perspective.
****
Distributed Bundle Adjustment with Block-Based Sparse Matrix Compression for Super Large Scale Datasets
Maoteng Zheng, Nengcheng Chen, Junfeng Zhu, Xiaoru Zeng, Huanbin Qiu, Yuyao Jiang, Xingyue Lu, Hao Qu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18152-18162
We propose a distributed bundle adjustment (DBA) method using the exact Levenberg-Marquardt (LM) algorithm for super large-scale datasets. Most of the existing methods partition the global map to small ones and conduct bundle adjustment in the submaps. In order to fit the parallel framework, they use approximate solutions instead of the LM algorithm. However, those methods often give sub-optimal results. Different from them, we utilize the exact LM algorithm to conduct global bundle adjustment where the formation of the reduced camera system (RCS) is actually parallelized and executed in a distributed way. To store the large RCS, we compress it with a block-based sparse matrix compression format (BSMC), which fully exploits its block feature. The BSMC format also enables the distributed storage and updating of the global RCS. The proposed method is extensively evaluated and compared with the state-of-the-art pipelines using both synthetic and real datasets. Preliminary results demonstrate the efficient memory usage and vast scalability of the proposed method compared with the baselines. For the first time, we conducted parallel bundle adjustment using LM algorithm on a real datasets with 1.18 million images and a synthetic dataset with 10 million images (about 500 times that of the state-of-the-art LM-based BA) on a distributed computing system.
****
SCANet: Scene Complexity Aware Network for Weakly-Supervised Video Moment Retrieval
Sunjae Yoon, Gwanhyeong Koo, Dahyun Kim, Chang D. Yoo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13576-13586
Video moment retrieval aims to localize moments in video corresponding to a given language query. To avoid the expensive cost of annotating the temporal moments, weakly-supervised VMR (wsVMR) systems have been studied. For such systems, generating a number of proposals as moment candidates and then selecting the most appropriate proposal has been a popular approach. These proposals are assumed to contain many distinguishable scenes in a video as candidates. However, existing proposals of wsVMR systems do not respect the varying numbers of scenes in each

video, where the proposals are heuristically determined irrespective of the video. We argue that the retrieval system should be able to counter the complexities caused by varying numbers of scenes in each video. To this end, we present a novel concept of a retrieval system referred to as Scene Complexity Aware Network (SCANet), which measures the `scene complexity' of multiple scenes in each video and generates adaptive proposals responding to variable complexities of scenes in each video. Experimental results on three retrieval benchmarks (i.e. Charades-STA, ActivityNet, TVR) achieve state-of-the-art performances and demonstrate the effectiveness of incorporating the scene complexity.
****
Neural Interactive Keypoint Detection
Jie Yang, Ailing Zeng, Feng Li, Shilong Liu, Ruimao Zhang, Lei Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15122-15132
This work proposes an end-to-end neural interactive keypoint detection framework named Click-Pose, which can significantly reduce more than 10 times labeling costs of 2D keypoint annotation compared with manual-only annotation. Click-Pose explores how user feedback can cooperate with a neural keypoint detector to correct the predicted keypoints in an interactive way for a faster and more effective annotation process. Specifically, we design the pose error modeling strategy that inputs the ground truth pose combined with four typical pose errors into the decoder and trains the model to reconstruct the correct poses, which enhances the self-correction ability of the model. Then, we attach an interactive human-feedback loop that allows receiving users' clicks to correct one or several predicted keypoints and iteratively utilizes the decoder to update all other keypoints with a minimum number of clicks (NoC) for efficient annotation. We validate Click-Pose in in-domain, out-of-domain scenes, and a new task of keypoint adaptation. For annotation, Click-Pose only needs 1.97 and 6.45 NoC@95 (at precision 95%) on COCO and Human-Art, reducing 31.4% and 36.3% efforts than the SOTA model (ViTPose) with manual correction, respectively. Besides, without user clicks, Click-Pose surpasses the previous end-to-end model by 1.4 AP on COCO and 3.0 AP on Human-Art.
****
Joint Implicit Neural Representation for High-fidelity and Compact Vector Fonts
Chia-Hao Chen, Ying-Tian Liu, Zhifei Zhang, Yuan-Chen Guo, Song-Hai Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5538-5548
Existing vector font generation approaches either struggle to preserve high-frequency corner details of the glyph or produce vector shapes that have redundant segments, which hinders their applications in practical scenarios. In this paper, we propose to learn vector fonts from pixelated font images utilizing a joint neural representation that consists of a signed distance field (SDF) and a probabilistic corner field (CF) to capture shape corner details. To achieve smooth shape interpolation on the learned shape manifold, we establish connections between the two fields for better alignment. We further design a vectorization process to extract high-quality and compact vector fonts from our joint neural representation. Experiments demonstrate that our method can generate more visually appealing vector fonts with a higher level of compactness compared to existing alternatives.
****
Spurious Features Everywhere - Large-Scale Detection of Harmful Spurious Features in ImageNet
Yannic Neuhaus, Maximilian Augustin, Valentyn Boreiko, Matthias Hein; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20235-20246
Benchmark performance of deep learning classifiers alone is not a reliable predictor for the performance of a deployed model. In particular, if the image classifier has picked up spurious features in the training data, its predictions can fail in unexpected ways. In this paper, we develop a framework that allows us to systematically identify spurious features in large datasets like ImageNet. It is

based on our neural PCA components and their visualization. Previous work on spurious features often operates in toy settings or requires costly pixel-wise annotations. In contrast, we work with ImageNet and validate our results by showing that presence of the harmful spurious feature of a class alone is sufficient to trigger the prediction of that class. We introduce the novel dataset "Spurious ImageNet" which allows to measure the reliance of any ImageNet classifier on harmful spurious features. Moreover, we introduce SpuFix as a simple mitigation method to reduce the dependence of any ImageNet classifier on previously identified harmful spurious features without requiring additional labels or retraining of the model. We provide code and data at https://github.com/YanNeu/spurious_imagenet.
****
Knowledge-Aware Prompt Tuning for Generalizable Vision-Language Models
Baoshuo Kan, Teng Wang, Wenpeng Lu, Xiantong Zhen, Weili Guan, Feng Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15670-15680
Pre-trained vision-language models, e.g., CLIP, working with manually designed prompts have demonstrated great effectiveness in transfer learning. Recently, learnable prompts achieve state-of-the-art performance, which however are prone to overfit to seen classes while failing to generalize to unseen classes. In this paper, we propose a Knowledge-Aware Prompt Tuning (KAPT) framework for vision-language models. Our approach takes the inspiration from human intelligence in which external knowledge is usually incorporated into recognizing novel categories of objects. Specifically, we design two complementary types of knowledge-aware prompts for the text encoder to leverage the distinctive characteristics of category-related external knowledge. The discrete prompt extracts the key information from descriptions of an object category, and the learned continuous prompt captures overall contexts. We further design an adaptation head for the visual encoder to aggregate salient attentive visual cues, which establishes discriminative and task-aware visual representations. We conduct extensive experiments on 11 widely-used benchmark datasets and the results verify the effectiveness in few-shot image classification, especially in generalizing to unseen categories. Compared with the state-of-the-art CoCoOp method, KAPT exhibits favorable performance and achieves an absolute gain of 3.22% on new classes and 2.57% in terms of harmonic mean.
****
Delicate Textured Mesh Recovery from NeRF via Adaptive Surface Refinement
Jiaxiang Tang, Hang Zhou, Xiaokang Chen, Tianshu Hu, Errui Ding, Jingdong Wang, Gang Zeng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17739-17749
Neural Radiance Fields (NeRF) have constituted a remarkable breakthrough in image-based 3D reconstruction.
However, their implicit volumetric representations differ significantly from the widely-adopted polygonal meshes and lack support from common 3D software and hardware, making their rendering and manipulation inefficient.
To overcome this limitation, we present a novel framework that generates textured surface meshes from images.
Our approach begins by efficiently initializing the geometry and view-dependency decomposed appearance with a NeRF.
Subsequently, a coarse mesh is extracted, and an iterative surface refinement algorithm is developed to adaptively adjust both vertex positions and face density based on re-projected rendering errors.
We jointly refine the appearance with geometry and bake it into texture images for real-time rendering.
Extensive experiments demonstrate that our method achieves superior mesh quality and competitive rendering quality.
****
Leveraging Inpainting for Single-Image Shadow Removal
Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, Song Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC

V), 2023, pp. 13055-13064
Fully-supervised shadow removal methods achieve the best restoration qualities on public datasets but still generate some shadow remnants. One of the reasons is the lack of large-scale shadow & shadow-free image pairs. Unsupervised methods can alleviate the issue but their restoration qualities are much lower than those of fully-supervised methods. In this work, we find that pretraining shadow removal networks on the image inpainting dataset can reduce the shadow remnants significantly: a naive encoder-decoder network gets competitive restoration quality w.r.t. the state-of-the-art methods via only 10% shadow & shadow-free image pairs. After analyzing networks with/without inpainting pretraining via the information stored in the weight (IIW), we find that inpainting pretraining improves restoration quality in non-shadow regions and enhances the generalization ability of networks significantly. Additionally, shadow removal fine-tuning enables networks to fill in the details of shadow regions. Inspired by these observations we formulate shadow removal as an adaptive fusion task that takes advantage of both shadow removal and image inpainting. Specifically, we develop an adaptive fusion network consisting of two encoders, an adaptive fusion block, and a decoder. The two encoders are responsible for extracting the features from the shadow image and the shadow-masked image respectively. The adaptive fusion block is responsible for combining these features in an adaptive manner. Finally, the decoder converts the adaptive fused features to the desired shadow-free result. The extensive experiments show that our method empowered with inpainting outperforms all state-of-the-art methods.
****

Neural Characteristic Function Learning for Conditional Image Generation

Shengxi Li, Jialu Zhang, Yifei Li, Mai Xu, Xin Deng, Li Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 7204-7214

The emergence of conditional generative adversarial networks (cGANs) has revolutionised the way we approach and control the generation, by means of adversarially learning joint distributions of data and auxiliary information. Despite the success, cGANs have been consistently put under scrutiny due to their ill-posed discrepancy measure between distributions, leading to mode collapse and instability problems in training. To address this issue, we propose a novel conditional characteristic function generative adversarial network (CCF-GAN) to reduce the discrepancy by the characteristic functions (CFs), which is able to learn accurate distance measure of joint distributions under theoretical soundness. More specifically, the difference between CFs is first proved to be complete and optimisation-friendly, for measuring the discrepancy of two joint distributions. To relieve the problem of curse of dimensionality in calculating CF difference, we propose to employ the neural network, namely neural CF (NCF), to efficiently minimise an upper bound of the difference. Based on the NCF, we establish the CCF-GAN framework to explicitly decompose CFs of joint distributions, which allows for learning the data distribution and auxiliary information with classified importance. The experimental results on synthetic and real-world datasets verify the superior performances of our CCF-GAN, on both the generation quality and stability.
****

Accurate 3D Face Reconstruction with Facial Component Tokens

Tianke Zhang, Xuangeng Chu, Yunfei Liu, Lijian Lin, Zhendong Yang, Zhengzhuo Xu, Chengkun Cao, Fei Yu, Changyin Zhou, Chun Yuan, Yu Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9033-9042

Accurately reconstructing 3D faces from monocular images and videos is crucial for various applications, such as digital avatar creation. However, the current deep learning-based methods face significant challenges in achieving accurate reconstruction with disentangled facial parameters and ensuring temporal stability in single-frame methods for 3D face tracking on video data. In this paper, we propose TokenFace, a transformer-based monocular 3D face reconstruction model. TokenFace uses separate tokens for different facial components to capture information about different facial parameters and employs temporal transformers to capture temporal information from video data. This design can naturally disentangle different facial components and is flexible to both 2D and 3D training data. Train

ed on hybrid 2D and 3D data, our model shows its power in accurately reconstruct
ing faces from images and producing stable results for video data. Experimental
results on popular benchmarks NoW and Stirling demonstrate that TokenFace achiev
es state-of-the-art performance, outperforming existing methods on all metrics b
y a large margin.
****
Holistic Label Correction for Noisy Multi-Label Classification
Xiaobo Xia, Jiankang Deng, Wei Bao, Yuxuan Du, Bo Han, Shiguang Shan, Tongliang
Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (IC
CV), 2023, pp. 1483-1493
Multi-label classification aims to learn classification models from instances as
sociated with multiple labels. It is pivotal to learn and utilize the label depe
ndence among multiple labels in multi-label classification. As a result of today
's big and complex data, noisy labels are inevitable, making it looming to targe
t multi-label classification with noisy labels. Although the importance of label
 dependence has been shown in multi-label classification with clean labels, it i
s challenging and hard to bring label dependence to the problem of multi-label c
lassification with noisy labels. The issues are, that we do not understand why l
abel dependence is helpful in the problem, and how to learn and utilize label de
pendence only using training data with noisy multiple labels. In this paper, we
bring label dependence to tackle the problem of multi-label classification with
noisy labels. Specifically, we first provide a high-level understanding of why l
abel dependence helps distinguish the examples with clean/noisy multiple labels.
 Benefiting from the memorization effect in handling noisy labels, a novel algor
ithm is then proposed to learn the label dependence by only employing training d
ata with noisy multiple labels, and utilize the learned dependence to help corre
ct noisy multiple labels to clean ones. We prove that the use of label dependenc
e could bring a higher success rate for recovering correct multiple labels. Empi
rical evaluations justify our claims and demonstrate the superiority of our algo
rithm.
****
Probabilistic Precision and Recall Towards Reliable Evaluation of Generative Mod
els
Dogyun Park, Suhyun Kim; Proceedings of the IEEE/CVF International Conference on
 Computer Vision (ICCV), 2023, pp. 20099-20109
Assessing the fidelity and diversity of the generative model is a difficult but
important issue for technological advancement. So, recent papers have introduced
 k-Nearest Neighbor (kNN) based precision-recall metrics to break down the stati
stical distance into fidelity and diversity. While they provide an intuitive met
hod, we thoroughly analyze these metrics and identify oversimplified assumptions
 and undesirable properties of kNN that result in unreliable evaluation, such as
 susceptibility to outliers and insensitivity to distributional changes. Thus, w
e propose novel metrics, P-precision and P-recall (PP&PR), based on a probabilis
tic approach that address the problems. Through extensive investigations on toy
experiments and state-of-the-art generative models, we show that our PP&PR provi
de more reliable estimates for comparing fidelity and diversity than the existin
g metrics. The codes are available at https://github.com/kdst-team/Probablistic_
precision_recall.
****
Deep Multitask Learning with Progressive Parameter Sharing
Haosen Shi, Shen Ren, Tianwei Zhang, Sinno Jialin Pan; Proceedings of the IEEE/C
VF International Conference on Computer Vision (ICCV), 2023, pp. 19924-19935
We propose a novel progressive parameter-sharing strategy (MPPS) in this paper f
or effectively training multitask learning models on diverse computer vision tas
ks simultaneously. Specifically, we propose to parameterize distributions for di
fferent tasks to control the sharings, based on the concept of Exclusive Capacit
y that we introduce. A scheduling mechanism following the concept of curriculum
learning
 is also designed to progressively change the sharing strategy to increase the l
evel of sharing during the learning process. We further propose a novel loss fun

ction to regularize the optimization of network parameters as well as the sharin
g probabilities of each neuron for each task. Our approach can be combined with
many state-of-the-art multitask learning solutions to achieve better joint task
performance.
 Comprehensive experiments show that it has competitive performance on three cha
llenging datasets (Multi-CIFAR100, NYUv2, and Cityscapes) using various convolut
ion neural network architectures.
****

Personalized Semantics Excitation for Federated Image Classification
Haifeng Xia, Kai Li, Zhengming Ding; Proceedings of the IEEE/CVF International C
onference on Computer Vision (ICCV), 2023, pp. 19301-19310
Federated learning casts a light on the collaboration of distributed local clien
ts with privacy protected to attain a more generic global model. However, signif
icant distribution shift in input/label space across different clients makes it
challenging to well generalize to all clients, which motivates personalized fede
rated learning (PFL). Existing PFL methods typically customize the local model b
y fine-tuning with limited local supervision and the global model regularizer, w
hich secures local specificity but risks ruining the global discriminative knowl
edge. In this paper, we propose a novel Personalized Semantics Excitation (PSE)
mechanism to breakthrough this limitation by exciting and fusing personalized se
mantics from the global model during local model customization. Specifically, PS
E explores channel-wise gradient differentiation across global and local models
to identify important low-level semantics mostly from convolutional layers which
 are embedded into the client-specific training. In addition, PSE deploys the co
llaboration of global and local models to enrich high-level feature representati
ons and facilitate the robustness of client classifier through a cross-model att
ention module. Extensive experiments and analysis on various image classificatio
n benchmarks demonstrate the effectiveness and advantage of our method over the
state-of-the-art PFL methods.
****

Unified Data-Free Compression: Pruning and Quantization without Fine-Tuning
Shipeng Bai, Jun Chen, Xintian Shen, Yixuan Qian, Yong Liu; Proceedings of the I
EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5876-5885
Structured pruning and quantization are promising approaches for reducing the in
ference time and memory footprint of neural networks. However, most existing met
hods require the original training dataset to fine-tune the model. This not only
 brings heavy resource consumption but also is not possible for applications wit
h sensitive or proprietary data due to privacy and security concerns. Therefore,
 a few data-free methods are proposed to address this problem, but they perform
data-free pruning and quantization separately, which does not explore the comple
mentarity of pruning and quantization. In this paper, we propose a novel framewo
rk named Unified Data-Free Compression(UDFC), which performs pruning and quantiz
ation simultaneously without any data and fine-tuning process. Specifically, UDF
C starts with the assumption that the partial information of a damaged(e.g., pru
ned or quantized) channel can be preserved by a linear combination of other chan
nels, and then derives the reconstruction form from the assumption to restore th
e information loss due to compression. Finally, we formulate the reconstruction
error between the original network and its compressed network, and theoretically
 deduce the closed-form solution. We evaluate the UDFC on the large-scale image
classification task and obtain significant improvements over various network arc
hitectures and compression methods. For example, we achieve a 20.54% accuracy im
provement on ImageNet dataset compared to SOTA method with 30% pruning ratio and
 6-bit quantization on ResNet-34.
****

SurroundOcc: Multi-camera 3D Occupancy Prediction for Autonomous Driving
Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, Jiwen Lu; Proceedings
of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21
729-21740
3D scene understanding plays a vital role in vision-based autonomous driving. Wh
ile most existing methods focus on 3D object detection, they have difficulty des

cribing real-world objects of arbitrary shapes and infinite classes. Towards a m
ore comprehensive perception of a 3D scene, in this paper, we propose a Surround
Occ method to predict the 3D occupancy with multi-camera images. We first extrac
t multi-scale features for each image and adopt spatial 2D-3D attention to lift
them to the 3D volume space. Then we apply 3D convolutions to progressively upsa
mple the volume features and impose supervision on multiple levels. To obtain de
nse occupancy prediction, we design a pipeline to generate dense occupancy groun
d truth without expansive occupancy annotations. Specifically, we fuse multi-fra
me LiDAR scans of dynamic objects and static scenes separately. Then we adopt Po
isson Reconstruction to fill the holes and voxelize the mesh to get dense occupa
ncy labels. Extensive experiments on nuScenes and SemanticKITTI datasets demonst
rate the superiority of our method. Code and dataset are available at https://gi
thub.com/weiyithu/SurroundOcc.
****
Temporal Enhanced Training of Multi-view 3D Object Detector via Historical Objec
t Prediction
Zhuofan Zong, Dongzhi Jiang, Guanglu Song, Zeyue Xue, Jingyong Su, Hongsheng Li,
Yu Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision
(ICCV), 2023, pp. 3781-3790
In this paper, we propose a new paradigm, named Historical Object Prediction (Ho
P) for multi-view 3D detection to leverage temporal information more effectively
. The HoP approach is straightforward: given the current timestamp t, we generat
e a pseudo Bird's-Eye View (BEV) feature of timestamp t-k from its adjacent fram
es and utilize this feature to predict the object set at timestamp t-k. Our appr
oach is motivated by the observation that enforcing the detector to capture both
the spatial location and temporal motion of objects occurring at historical tim
estamps can lead to more accurate BEV feature learning. First, we elaborately de
sign short-term and long-term temporal decoders, which can generate the pseudo B
EV feature for timestamp t-k without the involvement of its corresponding camera
images. Second, an additional object decoder is flexibly attached to predict th
e object targets using the generated pseudo BEV feature. Note that we only perfo
rm HoP during training, thus the proposed method does not introduce extra overhe
ads during inference. As a plug-and-play approach, HoP can be easily incorporate
d into state-of-the-art BEV detection frameworks, including BEVFormer and BEVDet
series. Furthermore, the auxiliary HoP approach is complementary to prevalent t
emporal modeling methods, leading to significant performance gains. Extensive ex
periments are conducted to evaluate the effectiveness of the proposed HoP on the
nuScenes dataset. We choose the representative methods, including BEVFormer and
BEVDet4D-Depth to evaluate our method. Surprisingly, HoP achieves 68.2% NDS and
61.6% mAP with ViT-L on nuScenes test, outperforming all the 3D object detector
s on the leaderboard by a large margin.
****
PARIS: Part-level Reconstruction and Motion Analysis for Articulated Objects
Jiayi Liu, Ali Mahdavi-Amiri, Manolis Savva; Proceedings of the IEEE/CVF Interna
tional Conference on Computer Vision (ICCV), 2023, pp. 352-363
We address the task of simultaneous part-level reconstruction and motion paramet
er estimation for articulated objects. Given two sets of multi-view images of an
object in two static articulation states, we decouple the movable part from the
static part and reconstruct shape and appearance while predicting the motion pa
rameters. To tackle this problem, we present PARIS: a self-supervised, end-to-en
d architecture that learns part-level implicit shape and appearance models and o
ptimizes motion parameters jointly without any 3D supervision, motion, or semant
ic annotation. Our experiments show that our method generalizes better across ob
ject categories, and outperforms baselines and prior work that are given 3D poin
t clouds as input. Our approach improves reconstruction relative to state-of-the
-art baselines with a Chamfer-L1 distance reduction of 3.94 (45.2%) for objects
and 26.79 (84.5%) for parts, and achieves 5% error rate for motion estimation ac
ross 10 object categories.
****
OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation

Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, Jianbing Shen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2761-2770

Referring video object segmentation (RVOS) aims at segmenting an object in a video following human instruction. Current state-of-the-art methods fall into an offline pattern, in which each clip independently interacts with text embedding for cross-modal understanding. They usually present that the offline pattern is necessary for RVOS, yet model limited temporal association within each clip. In this work, we break up the previous offline belief and propose a simple yet effective online model using explicit query propagation, named OnlineRefer. Specifically, our approach leverages target cues that gather semantic information and position prior to improve the accuracy and ease of referring predictions for the current frame. Furthermore, we generalize our online model into a semi-online framework to be compatible with video-based backbones. To show the effectiveness of our method, we evaluate it on four benchmarks, i.e., Refer-Youtube-VOS, Refer-DAVIS17, A2D-Sentences, and JHMDB-Sentences. Without bells and whistles, our OnlineRefer with a Swin-L backbone achieves 63.5 J&F and 64.8 J&F on Refer-Youtube-VOS and Refer-DAVIS17, outperforming all other offline methods. Our code is available at https://github.com/wudongming97/OnlineRefer.

****

Implicit Neural Representation for Cooperative Low-light Image Enhancement

Shuzhou Yang, Moxuan Ding, Yanmin Wu, Zihan Li, Jian Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12918-12927

The following three factors restrict the application of existing low-light image enhancement methods: unpredictable brightness degradation and noise, inherent gap between metric-favorable and visual-friendly versions, and the limited paired training data. To address these limitations, we propose an implicit Neural Representation method for Cooperative low-light image enhancement, dubbed NeRCo. It robustly recovers perceptual-friendly results in an unsupervised manner. Concretely, NeRCo unifies the diverse degradation factors of real-world scenes with a controllable fitting function, leading to better robustness. In addition, for the output results, we introduce semantic-orientated supervision with priors from the pre-trained vision-language model. Instead of merely following reference images, it encourages results to meet subjective expectations, finding more visual-friendly solutions. Further, to ease the reliance on paired data and reduce solution space, we develop a dual-closed-loop constrained enhancement module. It is trained cooperatively with other affiliated modules in a self-supervised manner. Finally, extensive experiments demonstrate the robustness and superior effectiveness of our proposed NeRCo. Our code is available at https://github.com/Ysz2022/NeRCo.

****

Environment Agnostic Representation for Visual Reinforcement Learning

Hyesong Choi, Hunsang Lee, Seongwon Jeong, Dongbo Min; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 263-273

Generalization capability of vision-based deep reinforcement learning (RL) is indispensable to deal with dynamic environment changes that exist in visual observations. The high-dimensional space of the visual input, however, imposes challenges in adapting an agent to unseen environments. In this work, we propose Environment Agnostic Reinforcement learning (EAR), which is a compact framework for domain generalization of the visual deep RL. Environment-agnostic features (EAFs) are extracted by leveraging three novel objectives based on feature factorization, reconstruction, and episode-aware state shifting, so that policy learning is accomplished only with vital features. EAR is a simple single-stage method with a low model complexity and a fast inference time, ensuring a high reproducibility, while attaining state-of-the-art performance in the DeepMind Control Suite and DrawerWorld benchmarks.

****

Deep Multiview Clustering by Contrasting Cluster Assignments

Jie Chen, Hua Mao, Wai Lok Woo, Xi Peng; Proceedings of the IEEE/CVF Internation

al Conference on Computer Vision (ICCV), 2023, pp. 16752-16761
Multiview clustering (MVC) aims to reveal the underlying structure of multiview data by categorizing data samples into clusters. Deep learning-based methods exhibit strong feature learning capabilities on large-scale datasets. For most existing deep MVC methods, exploring the invariant representations of multiple views is still an intractable problem. In this paper, we propose a cross-view contrastive learning (CVCL) method that learns view-invariant representations and produces clustering results by contrasting the cluster assignments among multiple views. Specifically, we first employ deep autoencoders to extract view-dependent features in the pretraining stage. Then, a cluster-level CVCL strategy is presented to explore consistent semantic label information among the multiple views in the fine-tuning stage. Thus, the proposed CVCL method is able to produce more discriminative cluster assignments by virtue of this learning strategy. Moreover, we provide a theoretical analysis of soft cluster assignment alignment. Extensive experimental results obtained on several datasets demonstrate that the proposed CVCL method outperforms several state-of-the-art approaches.
****
Mimic3D: Thriving 3D-Aware GANs via 3D-to-2D Imitation
Xingyu Chen, Yu Deng, Baoyuan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2338-2348
Generating images with both photorealism and multiview 3D consistency is crucial for 3D-aware GANs, yet existing methods struggle to achieve them simultaneously. Improving the photorealism via CNN-based 2D super-resolution can break the strict 3D consistency, while keeping the 3D consistency by learning high-resolution 3D representations for direct rendering often compromises image quality. In this paper, we propose a novel learning strategy, namely 3D-to-2D imitation, which enables a 3D-aware GAN to generate high-quality images while maintaining their strict 3D consistency, by letting the images synthesized by the generator's 3D rendering branch mimic those generated by its 2D super-resolution branch. We also introduce 3D-aware convolutions into the generator for better 3D representation learning, which further improves the image generation quality. With the above strategies, our method reaches FID scores of 5.4 and 4.3 on FFHQ and AFHQ-v2 Cats, respectively, at 512x512 resolution, largely outperforming existing 3D-aware GANs using direct 3D rendering and coming very close to the previous state-of-the-art method that leverages 2D super-resolution. Project website: https://seanchenxy.github.io/Mimic3DWeb.
****
Look at the Neighbor: Distortion-aware Unsupervised Domain Adaptation for Panoramic Semantic Segmentation
Xu Zheng, Tianbo Pan, Yunhao Luo, Lin Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18687-18698
Endeavors have been recently made to transfer knowledge from the labeled pinhole image domain to the unlabeled panoramic image domain via Unsupervised Domain Adaptation (UDA). The aim is to tackle the domain gaps caused by the style disparities and distortion problem of the non-uniformly distributed pixels of equirectangular projection (ERP). Previous works typically focus on transferring knowledge based on geometric priors with specially designed multi-branch network architectures. As a result, considerable computational costs are induced, and meanwhile, their generalization abilities are profoundly hindered by the variation of distortion among pixels. In this paper, we find that the pixels' neighborhood regions of the ERP indeed introduce less distortion. Intuitively, we propose a novel UDA framework that can effectively address the distortion problems for panoramic semantic segmentation. In comparison, our method is simpler, easier to implement, and more computationally efficient. Specifically, we propose distortion-aware attention (DA) capturing the neighboring pixel distribution without using any geometric constraints. Moreover, we propose a class-wise feature aggregation (CFA) module to iteratively update the feature representations with a memory bank. As such, the feature similarity between two domains can be consistently optimized. Extensive experiments show that our method achieves new state-of-the-art performance while remarkably reducing 80% parameters.

****
Rethinking Safe Semi-supervised Learning: Transferring the Open-set Problem to A Close-set One
Qiankun Ma, Jiyao Gao, Bo Zhan, Yunpeng Guo, Jiliu Zhou, Yan Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16370-16379
Conventional semi-supervised learning (SSL) lies in the close-set assumption that the labeled and unlabeled sets contain data with the same seen classes, called in-distribution (ID) data. In contrast, safe SSL investigates a more challenging open-set problem where unlabeled set may involve some out-of-distribution (OOD) data with unseen classes, which could harm the performance of SSL. When we are experimenting with the mainstream safe SSL methods, we have a surprising finding that all OOD data show a clear tendency to gather in the feature space. This inspires us to solve the safe SSL problem from a fresh perspective. Specifically, for a classification task with K seen classes, we utilize a prototype network not only to generate K prototypes of all seen classes, but also explicitly model an additional prototype for the OOD data, transferring the K-way classification on the open-set to the (K+1)-way on the close-set. In this way, the typical SSL techniques (e.g., consistency regularization and pseudo labeling) can be applied to tackle the safe SSL problem without additional consideration of OOD data processing like other safe SSL methods do. Particularly, considering the possible low-confidence pseudo labels, we further propose an iterative negative learning (INL) paradigm to enforce the network learning knowledge from complementary labels on wider classes, improving the network's classification performance. Extensive experiments on four benchmark datasets show that our approach remarkably lifts the performance on safe SSL and outperforms the state-of-the-art methods.
****
Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack
Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, Qi Alfred Chen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4412-4423
In autonomous driving (AD), accurate perception is indispensable to achieving safe and secure driving. Due to its safety-criticality, the security of AD perception has been widely studied. Among different attacks on AD perception, the physical adversarial object evasion attacks are especially severe. However, we find that all existing literature only evaluates their attack effect at the targeted AI component level but not at the system level, i.e., with the entire system semantics and context such as the full AD pipeline. Thereby, this raises a critical research question: can these existing researches effectively achieve system-level attack effects (e.g., traffic rule violations) in the real-world AD context? In this work, we conduct the first measurement study on whether and how effectively the existing designs can lead to system-level effects, especially for the STOP sign-evasion attacks due to their popularity and severity. Our evaluation results show that all the representative prior works cannot achieve any system-level effects. We observe two design limitations in the prior works: 1) physical model-inconsistent object size distribution in pixel sampling and 2) lack of vehicle plant model and AD system model consideration. Then, we propose SysAdv, a novel system-driven attack design in the AD context and our evaluation results show that the system-level effects can be significantly improved, i.e., the violation rate increases by around 70%.
****
ReLeaPS : Reinforcement Learning-based Illumination Planning for Generalized Photometric Stereo
Jun Hoong Chan, Bohan Yu, Heng Guo, Jieji Ren, Zongqing Lu, Boxin Shi; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9167-9175
Illumination planning in photometric stereo aims to find a balance between tween surface normal estimation accuracy and image capturing efficiency by selecting optimal light configurations. It depends on factors such as the unknown shape an

d general reflectance of the target object, global illumination, and the choice of photometric stereo backbones, which are too complex to be handled by existing methods based on handcrafted illumination planning rules. This paper proposes a learning-based illumination planning method that jointly considers these factors via integrating a neural network and a generalized image formation model. As it is impractical to supervise illumination planning due to the enormous search space for ground truth light configurations, we formulate illumination planning using reinforcement learning, which explores the light space in a photometric stereo-aware and reward-driven manner. Experiments on synthetic and real-world datasets demonstrate that photometric stereo under the 20-light configurations from our method is comparable to, or even surpasses that of using lights from all available directions.
****
Learning Foresightful Dense Visual Affordance for Deformable Object Manipulation
Ruihai Wu, Chuanruo Ning, Hao Dong; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10947-10956
Understanding and manipulating deformable objects (e.g., ropes and fabrics) is an essential yet challenging task with broad applications. Difficulties come from complex states and dynamics, diverse configurations and high-dimensional action space of deformable objects. Besides, the manipulation tasks usually require multiple steps to accomplish, and greedy policies may easily lead to local optimal states. Existing studies usually tackle this problem using reinforcement learning or imitating expert demonstrations, with limitations in modeling complex states or requiring hand-crafted expert policies. In this paper, we study deformable object manipulation using dense visual affordance, with generalization towards diverse states, and propose a novel kind of foresightful dense affordance, which avoids local optima by estimating states' values for long-term manipulation. We propose a framework for learning this representation, with novel designs such as multi-stage stable learning and efficient self-supervised data collection without experts. Experiments demonstrate the superiority of our proposed foresightful dense affordance.
****
Generalizable Neural Fields as Partially Observed Neural Processes
Jeffrey Gu, Kuan-Chieh Wang, Serena Yeung; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5330-5339
Neural fields, which represent signals as a function parameterized by a neural network, are a promising alternative to traditional discrete vector or grid-based representations. Compared to discrete representations, neural representations both scale well with increasing resolution, are continuous, and can be many-times differentiable. However, given a dataset of signals that we would like to represent, having to optimize a separate neural field for each signal is inefficient, and cannot capitalize on shared information or structures among signals. Existing generalization methods view this as a meta-learning problem and employ gradient-based meta-learning to learn an initialization which is then fine-tuned with test-time optimization, or learn hypernetworks to produce the weights of a neural field. We instead propose a new paradigm that views the large-scale training of neural representations as a part of a partially-observed neural process framework, and leverage neural process algorithms to solve this task. We demonstrate that this approach outperforms both state-of-the-art gradient-based meta-learning approaches and hypernetwork approaches.
****
CiteTracker: Correlating Image and Text for Visual Tracking
Xin Li, Yuqing Huang, Zhenyu He, Yaowei Wang, Huchuan Lu, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9974-9983
Existing visual tracking methods typically take an image patch as the reference of the target to perform tracking. However, a single image patch cannot provide a complete and precise concept of the target object as images are limited in their ability to abstract and can be ambiguous, which makes it difficult to track targets with drastic variations. In this paper, we propose the CiteTracking algor

ithm to enhance target modeling and inference in visual tracking by connecting images and text. Specifically, we develop a text generation module to convert the target image patch into a descriptive text containing its class and attribute information, providing a comprehensive reference point for the target. In addition, a dynamic description module is designed to adapt to target variations for more effective target representation. We then associate the target description and the search image using an attention-based correlation module to generate the correlated features for target state reference. Extensive experiments on five diverse datasets are conducted to evaluate the proposed algorithm and the favorable performance against the state-of-the-art methods demonstrates the effectiveness of the proposed tracking method. The source code and trained models will be made available to the public.
****
Adding Conditional Control to Text-to-Image Diffusion Models
Lvmin Zhang, Anyi Rao, Maneesh Agrawala; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3836-3847
We present ControlNet, a neural network architecture to add spatial conditioning controls to large, pretrained text-to-image diffusion models. ControlNet locks the production-ready large diffusion models, and reuses their deep and robust encoding layers pretrained with billions of images as a strong backbone to learn a diverse set of conditional controls. The neural architecture is connected with "zero convolutions" (zero-initialized convolution layers) that progressively grow the parameters from zero and ensure that no harmful noise could affect the finetuning. We test various conditioning controls, e.g., edges, depth, segmentation, human pose, etc., with Stable Diffusion, using single or multiple conditions, with or without prompts. We show that the training of ControlNets is robust with small (<50k) and large (>1m) datasets. Extensive results show that ControlNet may facilitate wider applications to control image diffusion models.
****
3D Instance Segmentation via Enhanced Spatial and Semantic Supervision
Salwa Al Khatib, Mohamed El Amine Boudjoghra, Jean Lahoud, Fahad Shahbaz Khan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 541-550
3D instance segmentation has recently garnered increased attention. Typical deep learning methods adopt point grouping schemes followed by hand-designed geometric clustering. Inspired by the success of transformers for various 3D tasks, newer hybrid approaches have utilized transformer decoders coupled with convolutional backbones that operate on voxelized scenes. However, due to the nature of sparse feature backbones, the extracted features provided to the transformer decoder are lacking in spatial understanding. Thus, such approaches often predict spatially separate objects as single instances. To this end, we introduce a novel approach for 3D point clouds instance segmentation that addresses the challenge of generating distinct instance masks for objects that share similar appearances but are spatially separated. Our method leverages spatial and semantic supervision with query refinement to improve the performance of hybrid 3D instance segmentation models. Specifically, we provide the transformer block with spatial features to facilitate differentiation between similar object queries and incorporate semantic supervision to enhance prediction accuracy based on object class. Our proposed approach outperforms existing methods on the validation sets of ScanNet V2 and ScanNet200 datasets, establishing a new state-of-the-art for this task.
****
Unleashing Text-to-Image Diffusion Models for Visual Perception
Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, Jiwen Lu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5729-5739
Diffusion models (DMs) have become the new trend of generative models and have demonstrated a powerful ability of conditional synthesis. Among those, text-to-image diffusion models pre-trained on large-scale image-text pairs are highly controllable by customizable prompts. Unlike the unconditional generative models that focus on low-level attributes and details, text-to-image diffusion models cont

ain more high-level knowledge thanks to the vision-language pre-training. In thi s paper, we propose VPD (Visual Perception with pre-trained Diffusion models), a new framework that exploits the semantic information of a pre-trained text-to-i mage diffusion model in visual perception tasks. Instead of using the pre-traine d denoising autoencoder in a diffusion-based pipeline, we simply use it as a bac kbone and aim to study how to take full advantage of the learned knowledge. Spec ifically, we prompt the denoising decoder with proper textual inputs and refine the text features with an adapter, leading to a better alignment to the pre-trai ned stage and making the visual contents interact with the text prompts. We also propose to utilize the cross-attention maps between the visual features and the text features to provide explicit guidance. Compared with other pre-training me thods, we show that vision-language pre-trained diffusion models can be faster a dapted to downstream visual perception tasks using the proposed VPD. Extensive e xperiments on semantic segmentation, referring image segmentation, and depth est imation demonstrate the effectiveness of our method. Notably, VPD attains 0.254 RMSE on NYUv2 depth estimation and 73.3% oIoU on RefCOCO-val referring image seg mentation, establishing new records on these two benchmarks. Code is available a t https://github.com/wl-zhao/VPD.
****
Iterative Superquadric Recomposition of 3D Objects from Multiple Views
Stephan Alaniz, Massimiliano Mancini, Zeynep Akata; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18013-18023
Humans are good at recomposing novel objects, i.e they can identify commonalitie s between unknown objects from general structure to finer detail, an ability dif ficult to replicate by machines. We propose a framework, ISCO, to recompose an o bject using 3D superquadrics as semantic parts directly from 2D views without tr aining a model that uses 3D supervision. To achieve this, we optimize the superq uadric parameters that compose a specific instance of the object, comparing its rendered 3D view and 2D image silhouette. Our ISCO framework iteratively adds ne w superquadrics wherever the reconstruction error is high, abstracting first coa rse regions and then finer details of the target object. With this simple coarse -to-fine inductive bias, ISCO provides consistent superquadrics for related obje ct parts, despite not having any semantic supervision. Since ISCO does not train any neural network, it is also inherently robust to out of distribution objects . Experiments show that, compared to recent single instance superquadrics recons truction approaches, ISCO provides consistently more accurate 3D reconstructions , even from images in the wild. Code available at https://github.com/Explainable ML/ISCO.
****
PHRIT: Parametric Hand Representation with Implicit Template
Zhisheng Huang, Yujin Chen, Di Kang, Jinlu Zhang, Zhigang Tu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14974-14 984
We propose PHRIT, a novel approach for parametric hand mesh modeling with an imp licit template that combines the advantages of both parametric meshes and implic it representations. Our method represents deformable hand shapes using signed di stance fields (SDFs) with part-based shape priors, utilizing a deformation field to execute the deformation. The model offers efficient high-fidelity hand recon struction by deforming the canonical template at infinite resolution. Additional ly, it is fully differentiable and can be easily used in hand modeling since it can be driven by the skeleton and shape latent codes. We evaluate PHRIT on multi ple downstream tasks, including skeleton-driven hand reconstruction, shapes from point clouds, and single-view 3D reconstruction, demonstrating that our approac h achieves realistic and immersive hand modeling with state-of-the-art performan ce.
****
BEVPlace: Learning LiDAR-based Place Recognition using Bird's Eye View Images
Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao, Junwei L i, Hui-Liang Shen; Proceedings of the IEEE/CVF International Conference on Compu ter Vision (ICCV), 2023, pp. 8700-8709

Place recognition is a key module for long-term SLAM systems. Current LiDAR-base d place recognition methods usually use representations of point clouds such as unordered points or range images. These methods achieve high recall rates of ret rieval, but their performance may degrade in the case of view variation or scene changes. In this work, we explore the potential of a different representation i n place recognition, i.e. bird's eye view (BEV) images. We validate that, withou t any delicate design, a simple ResNet trained on BEV images achieves comparable performance with the state-of-the-art place recognition methods in scenes of sl ight viewpoint changes. For more robust place recognition, we propose a rotation -invariant network called BEVPlace. We use group convolution to extract rotation -equivariant local features from the images and NetVLAD for global feature aggre gation. In addition, we observe that the distance between BEV features is correl ated with the geometry distance of point clouds. Based on the observation, we de velop a method to estimate the position of the query cloud, extending the usage of place recognition. The experiments conducted on large-scale public datasets s how that our method 1) achieves state-of-the-art performance in terms of recall rates, 2) is robust to view changes, 3) shows strong generalization ability, and 4) can estimate the positions of query point clouds. Source codes are publicly available at https://github.com/zjuluolun/BEVPlace.
****
Transferable Adversarial Attack for Both Vision Transformers and Convolutional N etworks via Momentum Integrated Gradients
Wenshuo Ma, Yidong Li, Xiaofeng Jia, Wei Xu; Proceedings of the IEEE/CVF Interna tional Conference on Computer Vision (ICCV), 2023, pp. 4630-4639
Visual Transformers (ViTs) and Convolutional Neural Networks (CNNs) are the two primary backbone structures extensively used in various vision tasks. Generating transferable adversarial examples for ViTs is difficult due to ViTs' superior r obustness, while transferring adversarial examples across ViTs and CNNs is even harder, since their structures and mechanisms for processing images are fundamen tally distinct. In this work, we propose a novel attack method named Momentum In tegrated Gradients (MIG), which not only attacks ViTs with high success rate, bu t also exhibits impressive transferability across ViTs and CNNs. Specifically, w e use integrated gradients rather than gradients to steer the generation of adve rsarial perturbations, inspired by the observation that integrated gradients of images demonstrate higher similarity across models in comparison to regular grad ients. Then we acquire the accumulated gradients by combining the integrated gra dients from previous iterations with the current ones in a momentum manner and u se their sign to modify the perturbations iteratively. We conduct extensive expe riments to demonstrate that adversarial examples obtained using MIG show stronge r transferability, resulting in significant improvements over state-of-the-art m ethods for both CNN and ViT models.
****
TrajPAC: Towards Robustness Verification of Pedestrian Trajectory Prediction Mod els
Liang Zhang, Nathaniel Xu, Pengfei Yang, Gaojie Jin, Cheng-Chao Huang, Lijun Zha ng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICC V), 2023, pp. 8327-8339
Robust pedestrian trajectory forecasting is crucial to developing safe autonomou s vehicles. Although previous works have studied adversarial robustness in the c ontext of trajectory forecasting, some significant issues remain unaddressed. In this work, we try to tackle these crucial problems. Firstly, the previous defin itions of robustness in trajectory prediction are ambiguous. We thus provide for mal definitions for two kinds of robustness, namely label robustness and pure ro bustness. Secondly, as previous works fail to consider robustness about all poin ts in a disturbance interval, we utilise a probably approximately correct (PAC) framework for robustness verification. Additionally, this framework can not only identify potential counterexamples, but also provides interpretable analyses of the original methods. Our approach is applied using a prototype tool named Traj PAC. With TrajPAC, we evaluate the robustness of four state-of-the-art trajector y prediction models -- Trajectron++, MemoNet, AgentFormer, and MID -- on traject

ories from five scenes of the ETH/UCY dataset and scenes of the Stanford Drone Dataset. Using our framework, we also experimentally study various factors that could influence robustness performance.
****
Adaptive Image Anonymization in the Context of Image Classification with Neural Networks
Nadiya Shvai, Arcadi Llanza Carmona, Amir Nakib; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5074-5083
Deep learning based methods have become the de-facto standard for various computer vision tasks. Nevertheless, they have repeatedly shown their vulnerability to various form of input perturbations such as pixels modification, region anonymization, etc. which are closely related to the adversarial attacks. This research particularly addresses the case of image anonymization, which is significantly important to preserve privacy and hence to secure digitized form of personal information from being exposed and potentially misused by different services that have captured it for various purposes. However, applying anonymization causes the classifier to provide different class decisions before and after applying it and therefore reduces the classifier's reliability and usability. In order to achieve a robust solution to this problem we propose a novel anonymization procedure that allows the existing classifiers to become class decision invariant on the anonymized images without any modification requires to apply on the classification models. We conduct numerous experiments on the popular ImageNet benchmark as well as on a large scale industrial toll classification problem's dataset. Obtained results confirm the efficiency and effectiveness of the proposed method as it obtained 0% rate of class decision change for both datasets compared to 15.95% on ImageNet and 0.18% on toll dataset obtained by applying the naive anonymization approaches. Moreover, it has shown a great potential to be applied to similar problems from different domains.
****
SiLK: Simple Learned Keypoints
Pierre Gleize, Weiyao Wang, Matt Feiszli; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22499-22508
Keypoint detection & descriptors are foundational technologies for computer vision tasks like image matching, 3D reconstruction and visual odometry. Hand-engineered methods like Harris corners, SIFT, and HOG descriptors have been used for decades; more recently, there has been a trend to introduce learning in an attempt to improve keypoint detectors. On inspection however, the results are difficult to interpret; recent learning-based methods employ a vast diversity of experimental setups and design choices: empirical results are often reported using different backbones, protocols, datasets, types of supervisions or tasks. Since these differences are often coupled together, it raises a natural question on what makes a good learned keypoint detector. In this work, we revisit the design of existing keypoint detectors by deconstructing their methodologies and identifying the key components. We re-design each component from first-principle and propose Simple Learned Keypoints (SiLK) that is fully-differentiable, lightweight, and flexible. Despite its simplicity, SiLK advances new state-of-the-art on Detection Repeatability and Homography Estimation tasks on HPatches and 3D Point-Cloud Registration task on ScanNet, and achieves competitive performance to state-of-the-art on camera pose estimation in 2022 Image Matching Challenge and ScanNet.
****
EfficientViT: Lightweight Multi-Scale Attention for High-Resolution Dense Prediction
Han Cai, Junyan Li, Muyan Hu, Chuang Gan, Song Han; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 17302-17313
High-resolution dense prediction enables many appealing real-world applications, such as computational photography, autonomous driving, etc. However, the vast computational cost makes deploying state-of-the-art high-resolution dense prediction models on hardware devices difficult. This work presents EfficientViT, a new family of high-resolution vision models with novel lightweight multi-scale attention. Unlike prior high-resolution dense prediction models that rely on heavy s

elf-attention, hardware-inefficient large-kernel convolution, or complicated top
ology structure to obtain good performances, our lightweight multi-scale attenti
on achieves a global receptive field and multi-scale learning (two critical feat
ures for high-resolution dense prediction) with only lightweight and hardware-ef
ficient operations. As such, EfficientViT delivers remarkable performance gains
over previous state-of-the-art high-resolution dense prediction models with sign
ificant speedup on diverse hardware platforms, including mobile CPU, edge GPU, a
nd cloud GPU. Without performance loss on Cityscapes, our EfficientViT provides
up to 8.8x and 3.8x GPU latency reduction over SegFormer and SegNeXt, respective
ly. For super-resolution, EfficientViT provides up to 6.4x speedup over Restorme
r while providing 0.11dB gain in PSNR.
****

Efficient Neural Supersampling on a Novel Gaming Dataset
Antoine Mercier, Ruan Erasmus, Yashesh Savani, Manik Dhingra, Fatih Porikli, Gui
llaume Berger; Proceedings of the IEEE/CVF International Conference on Computer
Vision (ICCV), 2023, pp. 296-306
Real-time rendering for video games has become increasingly challenging due to t
he need for higher resolutions, framerates and photorealism. Supersampling has e
merged as an effective solution to address this challenge. Our work introduces a
 novel neural algorithm for supersampling rendered content that is 4x more effic
ient than existing methods while maintaining the same level of accuracy. Additio
nally, we introduce a new dataset which provides auxiliary modalities such as mo
tion vectors and depth generated using graphics rendering features like viewport
 jittering and mipmap biasing at different resolutions. We believe that this dat
aset fills a gap in the current dataset landscape and can serve as a valuable re
source to help measure progress in the field and advance the state-of-the-art in
 super-resolution techniques for gaming content.
****

Rapid Adaptation in Online Continual Learning: Are We Evaluating It Right?
Hasan Abed Al Kader Hammoud, Ameya Prabhu, Ser-Nam Lim, Philip H.S. Torr, Adel B
ibi, Bernard Ghanem; Proceedings of the IEEE/CVF International Conference on Com
puter Vision (ICCV), 2023, pp. 18852-18861
We revisit the common practice of evaluating adaptation of Online Continual Lear
ning (OCL) algorithms through the metric of online accuracy, which measures the
accuracy of the model on the immediate next few samples. However, we show that t
his metric is unreliable, as even vacuous blind classifiers, which do not use in
put images for prediction, can achieve unrealistically high online accuracy by e
xploiting spurious label correlations in the data stream. Our study reveals that
 existing OCL algorithms can also achieve high online accuracy, but perform poor
ly in retaining useful information, suggesting that they unintentionally learn s
purious label correlations. To address this issue, we propose a novel metric for
 measuring adaptation based on the accuracy on the near-future samples, where sp
urious correlations are removed. We benchmark existing OCL approaches using our
proposed metric on large-scale datasets under various computational budgets and
find that better generalization can be achieved by retaining and reusing past se
en information. We believe that our proposed metric can aid in the development o
f truly adaptive OCL methods. We provide code to reproduce our results at https:
//github.com/drimpossible/EvalOCL.
****

Label-Efficient Online Continual Object Detection in Streaming Video
Jay Zhangjie Wu, David Junhao Zhang, Wynne Hsu, Mengmi Zhang, Mike Zheng Shou; P
roceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2
023, pp. 19246-19255
Humans can watch a continuous video stream and effortlessly perform continual ac
quisition and transfer of new knowledge with minimal supervision yet retaining p
reviously learnt experiences. In contrast, existing continual learning (CL) meth
ods require fully annotated labels to effectively learn from individual frames i
n a video stream. Here, we examine a more realistic and challenging problem--Lab
el-Efficient Online Continual Object Detection (LEOCOD) in streaming video. We p
ropose a plug-and-play module, Efficient-CLS, that can be easily inserted into a

nd improve existing continual learners for object detection in video streams with reduced data annotation costs and model retraining time. We show that our method has achieved significant improvement with minimal forgetting across all supervision levels on two challenging CL benchmarks for streaming real-world videos. Remarkably, with only 25% annotated video frames, our method still outperforms the base CL learners, which are trained with 100% annotations on all video frames. The data and source code will be publicly available at https://github.com/showlab/Efficient-CLS.
****
Learning Point Cloud Completion without Complete Point Clouds: A Pose-Aware Approach
Jihun Kim, Hyeokjun Kweon, Yunseo Yang, Kuk-Jin Yoon; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14203-14213
Point cloud completion is to restore complete 3D scenes and objects from incomplete observations or limited sensor data. Existing fully-supervised methods rely on paired datasets of incomplete and complete point clouds, which are labor-intensive to obtain. Unpaired methods have been proposed, but still require a set of complete point clouds as a reference. As a remedy, in this paper, we propose a novel point cloud completion framework without using any complete point cloud at all. Our main idea is to generate multiple incomplete point clouds of various poses and integrate them into a complete point cloud. We train our framework based on cycle consistency, to generate an incomplete point cloud such that 1) shares the same object as the input incomplete point cloud and 2) corresponds to an arbitrarily given pose. In addition, we devise a novel projection method conditioned by pose to gather visible features, from a volumetric feature extracted by an encoder. Extensive experiments demonstrate that the proposed method achieves comparable or better results than existing unpaired methods. Further, we show that our method also can be applied to real incomplete point clouds.
****
Frequency Guidance Matters in Few-Shot Learning
Hao Cheng, Siyuan Yang, Joey Tianyi Zhou, Lanqing Guo, Bihan Wen; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 11814-11824
Few-shot classification aims to learn a discriminative feature representation to recognize unseen classes with few labeled support samples. While most few-shot learning methods focus on exploiting the spatial information of image samples, frequency representation has also been proven essential in classification tasks. In this paper, we investigate the effect of different frequency components on the few-shot learning tasks. To enhance the performance and generalizability of few-shot methods, we propose a novel Frequency-Guided Few-shot Learning framework (dubbed FGFL), which leverages the task-specific frequency components to adaptively mask the corresponding image information, with a novel multi-level metric learning strategy including a triplet loss among original, masked and unmasked image as well as a contrastive loss between masked and original support and query sets to exploit more discriminative information. Extensive experiments on four benchmarks under several few-shot scenarios, i.e., standard, cross-dataset, cross-domain, and coarse-to-fine annotated classification, are conducted. Both qualitative and quantitative results show that our proposed FGFL scheme can attend to the class-discriminative frequency components, thus integrating those information towards more effective and generalizable few-shot learning.
****
Walking Your LiDOG: A Journey Through Multiple Domains for LiDAR Semantic Segmentation
Cristiano Saltori, Aljosa Osep, Elisa Ricci, Laura Leal-Taixé; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 196-206
The ability to deploy robots that can operate safely in diverse environments is crucial for developing embodied intelligent agents. As a community, we have made tremendous progress in within-domain LiDAR semantic segmentation. However, do these methods generalize across domains?
To answer this question, we design the first experimental setup for studying do

main generalization (DG) for LiDAR semantic segmentation (DG-LSS). Our results confirm a significant gap between methods, evaluated in a cross-domain setting: for example, a model trained on the source dataset (SemanticKITTI) obtains 26.53 mIoU on the target data, compared to 48.49 mIoU obtained by the model trained on the target domain (nuScenes).
To tackle this gap, we propose the first method specifically designed for DG-LSS, which obtains 34.88 mIoU on the target domain, outperforming all baselines. Our method augments a sparse-convolutional encoder-decoder 3D segmentation network with an additional, dense 2D convolutional decoder that learns to classify a birds-eye view of the point cloud. This simple auxiliary task encourages the 3D network to learn features that are robust to sensor placement shifts and resolution, and are transferable across domains. With this work, we aim to inspire the community to develop and evaluate future models in such cross-domain conditions.
****

Diverse Cotraining Makes Strong Semi-Supervised Segmentor
Yijiang Li, Xinjiang Wang, Lihe Yang, Litong Feng, Wayne Zhang, Ying Gao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16055-16067
Deep co-training has been introduced to semi-supervised segmentation and achieves impressive results, yet few studies have explored the working mechanism behind it. In this work, we revisit the core assumption that supports co-training: multiple compatible and conditionally independent views. By theoretically deriving the generalization upper bound, we prove the prediction similarity between two models negatively impacts the model's generalization ability. However, most current co-training models are tightly coupled together and violate this assumption. Such coupling leads to the homogenization of networks and confirmation bias which consequently limits the performance. To this end, we explore different dimensions of co-training and systematically increase the diversity from the aspects of input domains, different augmentations and model architectures to counteract homogenization. Our Diverse Co-training outperforms the state-of-the-art (SOTA) methods by a large margin across different evaluation protocols on the Pascal and Cityscapes. For example. we achieve the best mIoU of 76.2%, 77.7% and 80.2% on Pascal with only 92, 183 and 366 labeled images, surpassing the previous best results by more than 5%.
****

Spherical Space Feature Decomposition for Guided Depth Map Super-Resolution
Zixiang Zhao, Jiangshe Zhang, Xiang Gu, Chengli Tan, Shuang Xu, Yulun Zhang, Radu Timofte, Luc Van Gool; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12547-12558
Guided depth map super-resolution (GDSR), as a hot topic in multi-modal image processing, aims to upsample low-resolution (LR) depth maps with additional information involved in high-resolution (HR) RGB images from the same scene. The critical step of this task is to effectively extract domain-shared and domain-private RGB/depth features. In addition, three detailed issues, namely blurry edges, noisy surfaces, and over-transferred RGB texture, need to be addressed. In this paper, we propose the Spherical Space feature Decomposition Network (SSDNet) to solve the above issues. To better model cross-modality features, Restormer block-based RGB/depth encoders are employed for extracting local-global features. Then, the extracted features are mapped to the spherical space to complete the separation of private features and the alignment of shared features. Shared features of RGB are fused with the depth features to complete the GDSR task. Subsequently, a spherical contrast refinement (SCR) module is proposed to further address the detail issues. Patches that are classified according to imperfect categories are input into the SCR module, where the patch features are pulled closer to the ground truth and pushed away from the corresponding imperfect samples in the spherical feature space via contrastive learning. Extensive experiments demonstrate that our method can achieve state-of-the-art results on four test datasets, as well as successfully generalize to real-world scenes. The code is available at https://github.com/Zhaozixiang1228/GDSR-SSDNet.
****

Tiled Multiplane Images for Practical 3D Photography
Numair Khan, Lei Xiao, Douglas Lanman; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10454-10464
The task of synthesizing novel views from a single image has useful applications in virtual reality and mobile computing, and a number of approaches to the problem have been proposed in recent years. A Multiplane Image (MPI) estimates the scene as a stack of RGBA layers, and can model complex appearance effects, anti-alias depth errors and synthesize soft edges better than methods that use textured meshes or layered depth images. And unlike neural radiance fields, an MPI can be efficiently rendered on graphics hardware. However, MPIs are highly redundant and require a large number of depth layers to achieve plausible results. Based on the observation that the depth complexity in local image regions is lower than that over the entire image, we split an MPI into many small, tiled regions, each with only a few depth planes. We call this representation a Tiled Multiplane Image (TMPI). We propose a method for generating a TMPI with adaptive depth planes for single-view 3D photography in the wild. Our synthesized results are comparable to state-of-the-art single-view MPI methods while having lower computational overhead. each with only a few depth planes. We call this representation a Tiled Multiplane Image (TMPI). We propose a method for generating a TMPI with adaptive depth planes for single-view 3D photography in the wild. Our synthesized results are comparable to state-of-the-art single-view MPI methods while having lower computational overhead.
****
VQA-GNN: Reasoning with Multimodal Knowledge via Graph Neural Networks for Visual Question Answering
Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, Jure Leskovec; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21582-21592
Visual question answering (VQA) requires systems to perform concept-level reasoning by unifying unstructured (e.g., the context in question and answer; "QA context") and structured (e.g., knowledge graph for the QA context and scene; "concept graph") multimodal knowledge. Existing works typically combine a scene graph and a concept graph of the scene by connecting corresponding visual nodes and concept nodes, then incorporate the QA context representation to perform question answering. However, these methods only perform a unidirectional fusion from unstructured knowledge to structured knowledge, limiting their potential to capture joint reasoning over the heterogeneous modalities of knowledge. To perform more expressive reasoning, we propose VQA-GNN, a new VQA model that performs bidirectional fusion between unstructured and structured multimodal knowledge to obtain unified knowledge representations. Specifically, we inter-connect the scene graph and the concept graph through a super node that represents the QA context, and introduce a new multimodal GNN technique to perform inter-modal message passing for reasoning that mitigates representational gaps between modalities. On two challenging VQA tasks (VCR and GQA), our method outperforms strong baseline VQA methods by 3.2% on VCR (Q-AR) and 4.6% on GQA, suggesting its strength in performing concept-level reasoning. Ablation studies further demonstrate the efficacy of the bidirectional fusion and multimodal GNN method in unifying unstructured and structured multimodal knowledge.
****
Unmasked Teacher: Towards Training-Efficient Video Foundation Models
Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19948-19960
Video Foundation Models (VFMs) have received limited exploration due to high computational costs and data scarcity. Previous VFMs rely on Image Foundation Models (IFMs), which face challenges in transferring to the video domain. Although VideoMAE has trained a robust ViT from limited data, its low-level reconstruction poses convergence difficulties and conflicts with high-level cross-modal alignment. This paper proposes a training-efficient method for temporal-sensitive VFMs that integrates the benefits of existing methods. To increase data efficiency, w

e mask out most of the low-semantics video tokens, but selectively align the unmasked tokens with IFM, which serves as the UnMasked Teacher (UMT). By providing semantic guidance, our method enables faster convergence and multimodal friendliness. With a progressive pre-training framework, our model can handle various tasks including scene-related, temporal-related, and complex video-language understanding. Using only public sources for pre-training in 6 days on 32 A100 GPUs, our scratch-built ViT-L/16 achieves state-of-the-art performances on various video tasks.
****
Explore and Tell: Embodied Visual Captioning in 3D Environments
Anwen Hu, Shizhe Chen, Liang Zhang, Qin Jin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2482-2491
While current visual captioning models have achieved impressive performance, they often assume that the image is well-captured and provides a complete view of the scene. In real-world scenarios, however, a single image may not offer a good viewpoint, hindering fine-grained scene understanding. To overcome this limitation, we propose a novel task called Embodied Captioning, which equips visual captioning models with navigation capabilities, enabling them to actively explore the scene and reduce visual ambiguity from suboptimal viewpoints. Specifically, starting at a random viewpoint, an agent must navigate the environment to gather information from different viewpoints and generate a comprehensive paragraph describing all objects in the scene. To support this task, we build the ET-Cap dataset with Kubric simulator, consisting of 10K 3D scenes with cluttered objects and three annotated paragraphs per scene. We propose a Cascade Embodied Captioning model (CaBOT), which comprises of a navigator and a captioner, to tackle this task. The navigator predicts which actions to take in the environment, while the captioner generates a paragraph description based on the whole navigation trajectory. Extensive experiments demonstrate that our model outperforms other carefully designed baselines. Our dataset, codes and models are available at https://aim3-ruc.github.io/ExploreAndTell.
****
FastViT: A Fast Hybrid Vision Transformer Using Structural Reparameterization
Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, Anurag Ranjan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5785-5795
The recent amalgamation of transformer and convolutional designs has led to steady improvements in accuracy and efficiency of the models.
In this work, we introduce FastViT, a hybrid vision transformer architecture that obtains the state-of-the-art latency-accuracy trade-off. To this end, we introduce a novel token mixing operator, RepMixer, a building block of FastViT, that uses structural reparameterization to lower the memory access cost by removing skip-connections in the network. We further apply train-time overparametrization and large kernel convolutions to boost accuracy and empirically show that these choices have minimal effect on latency. We show that -- our model is 3.5x faster than CMT, a recent state-of-the-art hybrid transformer architecture, 4.9x faster than EfficientNet, and 1.9x faster than ConvNeXt on a mobile device for the same accuracy on the ImageNet dataset. At similar latency, our model obtains 4.2% better Top-1 accuracy on ImageNet than MobileOne. Our model consistently outperforms competing architectures across several tasks -- image classification, detection, segmentation and 3D mesh regression with significant improvement in latency on both a mobile device and a desktop GPU. Furthermore, our model is highly robust to out-of-distribution samples and corruptions, improving over competing robust models. Code and models are available at: https://github.com/apple/ml-fastvit
****
OFVL-MS: Once for Visual Localization across Multiple Indoor Scenes
Tao Xie, Kun Dai, Siyi Lu, Ke Wang, Zhiqiang Jiang, Jinghan Gao, Dedong Liu, Jie Xu, Lijun Zhao, Ruifeng Li; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5516-5526
In this work, we seek to predict camera poses across scenes with a multi-task le

arning manner, where we view the localization of each scene as a new task.
We propose OFVL-MS, a unified framework that dispenses with the traditional practice of training a model for each individual scene and relieves gradient conflict induced by optimizing multiple scenes collectively, enabling efficient storage yet precise visual localization for all scenes. Technically, in the forward pass of OFVL-MS, we design a layer-adaptive sharing policy with a learnable score for each layer to automatically determine whether the layer is shared or not. Such sharing policy empowers us to acquire task-shared parameters for a reduction of storage cost and task-specific parameters for learning scene-related features to alleviate gradient conflict. In the backward pass of OFVL-MS, we introduce a gradient normalization algorithm that homogenizes the gradient magnitude of the task-shared parameters so that all tasks converge at the same pace. Furthermore, a sparse penalty loss is applied on the learnable scores to facilitate parameter sharing for all tasks without performance degradation. We conduct comprehensive experiments on multiple benchmarks and our new released indoor dataset LIVL, showing that OFVL-MS families significantly outperform the state-of-the-arts with fewer parameters. We also verify that OFVL-MS can generalize to a new scene with much few parameters while gaining superior localization performance. The proposed dataset and evaluation code is available at https://github.com/mooncake199809/UFVL-Net.
****
HTML: Hybrid Temporal-scale Multimodal Learning Framework for Referring Video Object Segmentation
Mingfei Han, Yali Wang, Zhihui Li, Lina Yao, Xiaojun Chang, Yu Qiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13414-13423
Referring Video Object Segmentation (RVOS) is to segment the object instance from a given video, according to the textual description of this object. However, in the open world, the object descriptions are often diversified in contents and flexible in lengths. This leads to the key difficulty in RVOS, i.e., various descriptions of different ob- jects are corresponding to different temporal scales in the video, which is ignored by most existing approaches with single stride of frame sampling. To tackle this problem, we propose a concise Hybrid Temporal-scale Multimodal Learning (HTML) framework, which can effectively align lingual and visual features to discover core object semantics in the video, by learning multimodal interaction hierarchically from different temporal scales. More specifically, we introduce a novel inter-scale multimodal perception module, where the language queries dynamically interact with visual features across temporal scales. It can effectively reduce complex object confusion by passing video context among different scales. Finally, we conduct extensive experiments on the widely used benchmarks, including Ref- Youtube-VOS, Ref-DAVIS17, A2D-Sentences and JHMDB- Sentences, where our HTML achieves state-of-the-art performance on all these datasets.
****
SQAD: Automatic Smartphone Camera Quality Assessment and Benchmarking
Zilin Fang, Andrey Ignatov, Eduard Zamfir, Radu Timofte; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20532-20542
Smartphone photography is becoming increasingly popular, but fitting high-performing camera systems within the given space limitations remains a challenge for manufacturers. As a result, powerful mobile camera systems are in high demand. Despite recent progress in computer vision, camera system quality assessment remains a tedious and manual process. In this paper, we present the Smartphone Camera Quality Assessment Dataset (SQAD), which includes natural images captured by 29 devices. SQAD defines camera system quality based on six widely accepted criteria: resolution, color accuracy, noise level, dynamic range, Point Spread Function, and aliasing. Built on thorough examinations in a controlled laboratory environment, SQAD provides objective metrics for quality assessment, overcoming previous subjective opinion scores. Moreover, we introduce the task of automatic camera quality assessment and train deep learning-based models on the collected data to perform a precise quality prediction for arbitrary photos. The dataset, code

s and pre-trained models are released at https://github.com/aiff22/SQAD.
****
PointDC: Unsupervised Semantic Segmentation of 3D Point Clouds via Cross-Modal Distillation and Super-Voxel Clustering
Zisheng Chen, Hongbin Xu, Weitao Chen, Zhipeng Zhou, Haihong Xiao, Baigui Sun, Xuansong Xie, Wenxiong kang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14290-14299
Semantic segmentation of point clouds usually requires exhausting efforts of human annotations, hence it attracts wide attention to a challenging topic of learning from unlabeled or weaker form of annotations. In this paper, we take the first attempt for fully unsupervised semantic segmentation of point clouds, which aims to delineate semantically meaningful objects without any form of annotations. Previous works of unsupervised pipeline on 2D images fails in this task of point clouds, due to: 1) Clustering Ambiguity caused by limited magnitude of data and imbalanced class distribution; 2) Irregularity Ambiguity caused by the irregular sparsity of point cloud. Therefore, we propose a novel framework, PointDC, which is comprised of two steps that handles the aforementioned problems respectively: Cross-Modal Distillation (CVD) and Super-Voxel Clustering (SVC). In the first stage of CVD, multi-view visual features are back-projected to the 3D space and aggregated to a unified point feature to distill the training of the point representation. In the second stage of SVC, the point features are aggregated to super-voxels and then fed to the iterative clustering process for excavating semantic classes. PointDC yields a significant improvement over the prior state-of-the-art unsupervised methods, on both the ScanNet v2 (+18.4 mIOU) and S3DIS (+11.5 mIOU) semantic segmentation benchmarks.
****
MV-Map: Offboard HD-Map Generation with Multi-view Consistency
Ziyang Xie, Ziqi Pang, Yu-Xiong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8658-8668
While bird's-eye-view (BEV) perception models can be useful for building high-definition maps (HD-Maps) with less human labor, their results are often unreliable and demonstrate noticeable inconsistencies in the predicted HD-Maps from different viewpoints. This is because BEV perception is typically set up in an "onboard" manner, which restricts the computation and consequently prevents algorithms from reasoning multiple views simultaneously. This paper overcomes these limitations and advocates a more practical "offboard" HD-Map generation setup that removes the computation constraints, based on the fact that HD-Maps are commonly reusable infrastructures built offline in data centers.
To this end, we propose a novel offboard pipeline called MV-Map that capitalizes multi-view consistency and can handle an arbitrary number of frames with the key desgin of a "region-centric" framework. In MV-Map, the target HD-Maps are created by aggregating all the frames of onboard predictions, weighted by the confidence scores assigned by an "uncertainty network." To further enhance multi-view consistency, we augment the uncertainty network with the global 3D structure optimized by a voxelized neural radiance field (Voxel-NeRF). Extensive experiments on nuScenes show that our MV-Map significantly improves the quality of HD-Maps, further highlighting the importance of offboard methods for HD-Map generation.
****
Multi-view Self-supervised Disentanglement for General Image Denoising
Hao Chen, Chenyuan Qu, Yu Zhang, Chen Chen, Jianbo Jiao; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12281-12291
With its significant performance improvements, the deep learning paradigm has become a standard tool for modern image denoisers. While promising performance has been shown on seen noise distributions, existing approaches often suffer from generalisation to unseen noise types or general and real noise. It is understandable as the model is designed to learn paired mapping (e.g. from a noisy image to its clean version). In this paper, we instead propose to learn to disentangle the noisy image, under the intuitive assumption that different corrupted versions of the same clean image share a common latent space. A self-supervised learning framework is proposed to achieve the goal, without looking at the latent clean

image. By taking two different corrupted versions of the same image as input, the proposed Multi-view Self-supervised Disentanglement (MeD) approach learns to disentangle the latent clean features from the corruptions and recover the clean image consequently. Extensive experimental analysis on both synthetic and real noise shows the superiority of the proposed method over prior self-supervised approaches, especially on unseen novel noise types. On real noise, the proposed method even outperforms its supervised counterparts by over 3dB.
****
Inter-Realization Channels: Unsupervised Anomaly Detection Beyond One-Class Classification
Declan McIntosh, Alexandra Branzan Albu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6285-6295
Unsupervised anomaly detection and localization in images is a challenging problem, leading previous methods to attempt an easier supervised one-class classification formalization. Assuming training images to be realizations of the underlying image distribution, it follows that nominal patches from these realizations will be well associated between and represented across realizations. From this, we propose Inter-Realization Channels (InReaCh), a fully unsupervised method of detecting and localizing anomalies. InReaCh extracts high-confidence nominal patches from training data by associating them between realizations into channels, only considering channels with high spans and low spread as nominal. We then create our nominal model from the patches of these channels to test new patches against. InReaCh extracts nominal patches from the MVTec AD dataset with 99.9% precision, then archives 0.968 AUROC in localization and 0.923 AUROC in detection with corrupted training data, competitive with current state-of-the-art supervised one-class classification methods. We test our model up to 40% of training data containing anomalies with negligibly affected performance. The shift to fully unsupervised training simplifies dataset creation and broadens possible applications.
****
Multi-Event Video-Text Retrieval
Gengyuan Zhang, Jisen Ren, Jindong Gu, Volker Tresp; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 22113-22123
Video-Text Retrieval (VTR) is a crucial multi-modal task in an era of massive video-text data on the Internet.
A plethora of work characterized by using a two-stream Vision-Language model architecture that learns a joint representation of video-text pairs has become a prominent approach for the VTR task.
However, these models operate under the assumption of bijective video-text correspondences and neglect a more practical scenario where video content usually encompasses multiple events, while texts like user queries or webpage metadata tend to be specific and correspond to single events.
This establishes a gap between the previous training objective and real-world applications, leading to the potential performance degradation of earlier models during inference.
In this study, we introduce the Multi-event Video-Text Retrieval (MeVTR) task, addressing scenarios in which each video contains multiple different events, as a niche scenario of the conventional Video-Text Retrieval Task. We present a simple model, Me-Retriever, which incorporates key event video representation and a new MeVTR loss for the MeVTR task. Comprehensive experiments show that this straightforward framework outperforms other models in the Video-to-Text and Text-to-Video tasks, effectively establishing a robust baseline for the MeVTR task. We believe this work serves as a strong foundation for future studies.
Code is available at https://github.com/gengyuanmax/MeVTR.
****
SHERF: Generalizable Human NeRF from a Single Image
Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, Ziwei Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9352-9364
Existing Human NeRF methods for reconstructing 3D humans typically rely on multi

ple 2D images from multi-view cameras or monocular videos captured from fixed ca mera views. However, in real-world scenarios, human images are often captured fr om random camera angles, presenting challenges for high-quality 3D human reconst ruction. In this paper, we propose SHERF, the first generalizable Human NeRF mod el for recovering animatable 3D humans from a single input image. SHERF extracts and encodes 3D human representations in canonical space, enabling rendering and animation from free views and poses. To achieve high-fidelity novel view and po se synthesis, the encoded 3D human representations should capture both global ap pearance and local fine-grained textures. To this end, we propose a bank of 3D-a ware hierarchical features, including global, point-level, and pixel-aligned fea tures, to facilitate informative encoding. Global features enhance the informati on extracted from the single input image and complement the information missing from the partial 2D observation. Point-level features provide strong clues of 3D human structure, while pixel-aligned features preserve more fine-grained detail s. To effectively integrate the 3D-aware hierarchical feature bank, we design a feature fusion transformer. Extensive experiments on THuman, RenderPeople, ZJU_M oCap, and HuMMan datasets demonstrate that SHERF achieves state-of-the-art perfo rmance, with better generalizability for novel view and pose synthesis.
****
MVPSNet: Fast Generalizable Multi-view Photometric Stereo
Dongxu Zhao, Daniel Lichy, Pierre-Nicolas Perrin, Jan-Michael Frahm, Soumyadip S engupta; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12525-12536
We propose a fast and generalizable solution to Multiview Photometric Stereo (MV PS), called MVPSNet. The key to our approach is a feature extraction network tha t effectively combines images from the same view captured under multiple lightin g conditions to extract geometric features from shading cues for stereo matching . We demonstrate these features, termed 'Light Aggregated Feature Maps' (LAFM), are effective for feature matching even in textureless regions, where traditiona l multi-view stereo methods often fail. Our method produces similar reconstructi on results to PS-NeRF, a state-of-the-art MVPS method that optimizes a neural ne twork per-scene, while being 411x faster (105 seconds vs. 12 hours) in inference . Additionally, we introduce a new synthetic dataset for MVPS, sMVPS, which is s hown to be effective for training a generalizable MVPS method.
****
High Quality Entity Segmentation
Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia , Zhe Lin, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 4047-4056
Dense image segmentation tasks (e.g., semantic, panop tic) are useful for image editing, but existing methods can hardly generalize well in an in-the-wild setti ng where there are unrestricted image domains, classes, and image reso lution & quality variations. Motivated by these observa tions, we construct a new entity segmentation dataset, with a strong focus on high-quality dense segmentation in the wild. The dataset contains images spanning diverse image domains and entitie s, along with plent(ful high-resolution images and high-quality mask annotations for training and testing. Given the high-quality and -resolution nature of the dataset, we propose CropFormer which is designed to tackle the intractability of instance-level segmentation on high-resolution images. It improves mask predict ion by fusing high-res image crops that provides more fine grained image details and the full image. CropFormer is the first query-based Tran. former architectu re that can ef fectively fuse mask predictions from multiple image views, by lea rning queries that effectively associate the same en tities across the full imag e and its crop. With CropFormer, we achieve a significant AP gain of 1.9 on the challenging entity segmentation task. Furthermore, CropFormer con sistently impr oves the accuracy of traditional segmentation tasks and datasets. The dataset an d code are released at http://luqi.info/entityv2.github.iol
****
CoTDet: Affordance Knowledge Prompting for Task Driven Object Detection
Jiajin Tang, Ge Zheng, Jingyi Yu, Sibei Yang; Proceedings of the IEEE/CVF Intern

ational Conference on Computer Vision (ICCV), 2023, pp. 3068-3078
Task driven object detection aims to detect object instances suitable for afford
ing a task in an image. Its challenge lies in object categories available for th
e task being too diverse to be limited to a closed set of object vocabulary for
traditional object detection. Simply mapping categories and visual features of c
ommon objects to the task cannot address the challenge. In this paper, we propos
e to explore fundamental affordances rather than object categories, i.e., common
 attributes that enable different objects to accomplish the same task. Moreover,
 we propose a novel multi-level chain-of-thought prompting (MLCoT) to extract th
e affordance knowledge from large language models, which contains multi-level re
asoning steps from task to object examples to essential visual attributes with r
ationales. Furthermore, to fully exploit knowledge to benefit object recognition
 and localization, we propose a knowledge-conditional detection framework, namel
y CoTDet. It conditions the detector from the knowledge to generate object queri
es and regress boxes. Experimental results demonstrate that our CoTDet outperfor
ms state-of-the-art methods consistently and significantly (+15.6 box AP and +14
.8 mask AP) and can generate rationales for why objects are detected to afford t
he task.
****
You Never Get a Second Chance To Make a Good First Impression: Seeding Active Le
arning for 3D Semantic Segmentation
Nermin Samet, Oriane Siméoni, Gilles Puy, Georgy Ponimatkin, Renaud Marlet, Vinc
ent Lepetit; Proceedings of the IEEE/CVF International Conference on Computer Vi
sion (ICCV), 2023, pp. 18445-18457
We propose SeedAL, a method to seed active learning for efficient annotation of
3D point clouds for semantic segmentation. Active Learning (AL) iteratively sele
cts relevant data fractions to annotate within a given budget, but requires a fi
rst fraction of the dataset (a 'seed') to be already annotated to estimate the b
enefit of annotating other data fractions. We first show that the choice of the
seed can significantly affect the performance of many AL methods. We then propos
e a method for automatically constructing a seed that will ensure good performan
ce for AL. Assuming that images of the point clouds are available, which is comm
on, our method relies on powerful unsupervised image features to measure the div
ersity of the point clouds. It selects the point clouds for the seed by optimizi
ng the diversity under an annotation budget, which can be done by solving a line
ar optimization problem. Our experiments demonstrate the effectiveness of our ap
proach compared to random seeding and existing methods on both the S3DIS and Sem
anticKitti datasets. Code is available at https://github.com/nerminsamet/seedal.
****
Scalable Multi-Temporal Remote Sensing Change Data Generation via Simulating Sto
chastic Change Process
Zhuo Zheng, Shiqi Tian, Ailong Ma, Liangpei Zhang, Yanfei Zhong; Proceedings of
the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21818
-21827
Understanding the temporal dynamics of Earth's surface is a mission of multi-tem
poral remote sensing image analysis, significantly promoted by deep vision model
s with its fuel---labeled multi-temporal images. However, collecting, preprocess
ing, and annotating multi-temporal remote sensing images at scale is non-trivial
 since it is expensive and knowledge-intensive. In this paper, we present a scal
able multi-temporal remote sensing change data generator via generative modeling
, which is cheap and automatic, alleviating these problems. Our main idea is to
simulate a stochastic change process over time. We consider the stochastic chang
e process as a probabilistic semantic state transition, namely generative probab
ilistic change model (GPCM), which decouples the complex simulation problem into
 two more trackable sub-problems, i.e., change event simulation and semantic cha
nge synthesis. To solve these two problems, we present the change generator (Cha
ngen), a GAN-based GPCM, enabling controllable object change data generation, in
cluding customizable object property, and change event. The extensive experiment
s suggest that our Changen has superior generation capability, and the change de
tectors with Changen pre-training exhibit excellent transferability to real-worl

d change datasets.
****
Human from Blur: Human Pose Tracking from Blurry Images
Yiming Zhao, Denys Rozumnyi, Jie Song, Otmar Hilliges, Marc Pollefeys, Martin R. Oswald; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 14905-14915
We propose a method to estimate 3D human poses from substantially blurred images. The key idea is to tackle the inverse problem of image deblurring by modeling the forward problem with a 3D human model, a texture map, and a sequence of poses to describe human motion. The blurring process is then modeled by a temporal image aggregation step. Using a differentiable renderer, we can solve the inverse problem by backpropagating the pixel-wise reprojection error to recover the best human motion representation that explains a single or multiple input images. Since the image reconstruction loss alone is insufficient, we present additional regularization terms. To the best of our knowledge, we present the first method to tackle this problem. Our method consistently outperforms other methods on significantly blurry inputs since they lack one or multiple key functionalities that our method unifies, i.e. image deblurring with sub-frame accuracy and explicit 3D modeling of non-rigid human motion.
****
NerfAcc: Efficient Sampling Accelerates NeRFs
Ruilong Li, Hang Gao, Matthew Tancik, Angjoo Kanazawa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18537-18546
Optimizing and rendering Neural Radiance Fields is computationally expensive due to the vast number of samples required by volume rendering. Recent works have included alternative sampling approaches to help accelerate their methods, however, they are often not the focus of the work. In this paper, we investigate and compare multiple sampling approaches and demonstrate that improved sampling is generally applicable across NeRF variants under an unified concept of transmittance estimator. To facilitate future experiments, we develop NerfAcc, a Python toolbox that provides flexible APIs for incorporating advanced sampling methods into NeRF related methods. We demonstrate its flexibility by showing that it can reduce the training time of several recent NeRF methods by 1.5x to 20x with minimal modifications to the existing codebase. Additionally, highly customized NeRFs, such as Instant-NGP, can be implemented in native PyTorch using NerfAcc. Our code are open-sourced at https://www.nerfacc.com.
****
A2Q: Accumulator-Aware Quantization with Guaranteed Overflow Avoidance
Ian Colbert, Alessandro Pappalardo, Jakoba Petri-Koenig; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 16989-16998
We present accumulator-aware quantization (A2Q), a novel weight quantization method designed to train quantized neural networks (QNNs) to avoid overflow when using low-precision accumulators during inference. A2Q introduces a unique formulation inspired by weight normalization that constrains the L1-norm of model weights according to accumulator bit width bounds that we derive. Thus, in training QNNs for low-precision accumulation, A2Q also inherently promotes unstructured weight sparsity to guarantee overflow avoidance. We apply our method to deep learning-based computer vision tasks to show that A2Q can train QNNs for low-precision accumulators while maintaining model accuracy competitive with a floating-point baseline. In our evaluations, we consider the impact of A2Q on both general-purpose platforms and programmable hardware. However, we primarily target model deployment on FPGAs because they can be programmed to fully exploit custom accumulator bit widths. Our experimentation shows accumulator bit width significantly impacts the resource efficiency of FPGA-based accelerators. On average across our benchmarks, A2Q offers up to a 2.3x reduction in resource utilization over 32-bit accumulator counterparts with 99.2% of the floating-point model accuracy.
****
Uni-3D: A Universal Model for Panoptic 3D Scene Reconstruction
Xiang Zhang, Zeyuan Chen, Fangyin Wei, Zhuowen Tu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9256-9266

Performing holistic 3D scene understanding from a single-view observation, invol ving generating instance shapes and 3D scene segmentation, is a long-standing ch allenge. Prevailing works either focus only on geometry or segmentation, or mode l the task in two folds by separate modules, whose results are merged later to f orm the final prediction. Inspired by recent advances in 2D vision that unify im age segmentation and detection by Transformer-based models, we present Uni-3D, a holistic 3D scene parsing/reconstruction system for a single RGB image. Uni-3D features a universal model with query-based representations for predicting segme nts of both object instances and scene layout. In Uni-3D, we also introduce a si ngle Transformer for 2D depth-aware panoptic segmentation, which offers queries that serve as strong shape priors in 3D. Uni-3D seamlessly integrates 2D and 3D in its architecture and it outperforms previous methods significantly.
****
ARNOLD: A Benchmark for Language-Grounded Task Learning with Continuous States i n Realistic 3D Scenes
Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, Baoxiong Jia, Siyuan Huang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20483-20495
Understanding the continuous states of objects is essential for task learning an d planning in the real world. However, most existing task learning benchmarks as sume discrete (e.g., binary) object states, which poses challenges for learning complex tasks and transferring learned policy from the simulated environment to the real world. Furthermore, the robot's ability to follow human instructions ba sed on grounding the actions and states is limited. To tackle these challenges, we present ARNOLD, a benchmark that evaluates language-grounded task learning wi th continuous states in realistic 3D scenes. ARNOLD consists of 8 language-condi tioned tasks that involve understanding object states and learning policies for continuous goals. To promote language-instructed learning, we provide expert dem onstrations with template-generated language descriptions. We assess task perfor mance by utilizing the latest language-conditioned policy learning models. Our r esults indicate that current models for language-conditioned manipulations conti nue to experience significant challenges when it comes to novel goal-state gener alizations, scene generalizations, and object generalizations. These findings hi ghlight the need to develop new algorithms to address this gap and underscore th e potential for further research in this area.
****
Full-Body Articulated Human-Object Interaction
Nan Jiang, Tengyu Liu, Zhexuan Cao, Jieming Cui, Zhiyuan Zhang, Yixin Chen, He W ang, Yixin Zhu, Siyuan Huang; Proceedings of the IEEE/CVF International Conferen ce on Computer Vision (ICCV), 2023, pp. 9365-9376
Fine-grained capture of 3D Human-Object Interactions (HOIs) boosts human activit y understanding and facilitates various downstream visual tasks. Prior models mo stly assume that humans interact with rigid objects using only a few body parts, limiting their scope. In this paper, we address the challenging problem of Full -Body Articulated Human-Object Interaction (f-AHOI), wherein the whole human bod ies interact with articulated objects, whose parts are connected by movable join ts. We present Capturing Human and Articulated-object InteRactionS (CHAIRS), a l arge-scale motion-captured f-AHOI dataset, consisting of 17.3 hours of versatile interactions between 46 participants and 81 articulated and rigid sittable obje cts. CHAIRS provides 3D meshes of both humans and articulated objects during the entire interactive process, as well as realistic and physically plausible full- body interactions. We show the value of CHAIRS with object pose estimation. By l earning the geometrical relationships in HOI, we devise the first model that lev erages human pose estimation to tackle the articulated object pose/shape estimat ion during whole-body interactions. Given an image and an estimated human pose, our model reconstructs the object pose/shape and optimizes the reconstruction ac cording to a learned interaction prior. Under two evaluation settings, our model significantly outperforms baselines. We further demonstrate the value of CHAIRS with a downstream task of generating interacting human poses conditioned on art

iculated objects. We hope CHAIRS will promote the community towards finer-grained interaction understanding. Data/code will be made publicly available.
****
FeatureNeRF: Learning Generalizable NeRFs by Distilling Foundation Models
Jianglong Ye, Naiyan Wang, Xiaolong Wang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8962-8973
Recent works on generalizable NeRFs have shown promising results on novel view synthesis from single or few images. However, such models have rarely been applied on other downstream tasks beyond synthesis such as semantic understanding and parsing. In this paper, we propose a novel framework named FeatureNeRF to learn generalizable NeRFs by distilling pre-trained vision foundation models (e.g., DINO, Latent Diffusion). FeatureNeRF leverages 2D pre-trained foundation models to 3D space via neural rendering, and then extract deep features for 3D query points from NeRF MLPs. Consequently, it allows to map 2D images to continuous 3D semantic feature volumes, which can be used for various downstream tasks. We evaluate FeatureNeRF on tasks of 2D/3D semantic keypoint transfer and 2D/3D object part segmentation. Our extensive experiments demonstrate the effectiveness of FeatureNeRF as a generalizable 3D semantic feature extractor.
****
SRFormer: Permuted Self-Attention for Single Image Super-Resolution
Yupeng Zhou, Zhen Li, Chun-Le Guo, Song Bai, Ming-Ming Cheng, Qibin Hou; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 12780-12791
Previous works have shown that increasing the window size for Transformer-based image super-resolution models (e.g., SwinIR) can significantly improve the model performance but the computation overhead is also considerable. In this paper, we present SRFormer, a simple but novel method that can enjoy the benefit of large window self-attention but introduces even less computational burden. The core of our SRFormer is the permuted self-attention(PSA), which strikes an appropriate balance between the channel and spatial information for self-attention. Our PSA is simple and can be easily applied to existing super-resolution networks based on window self-attention. Without any bells and whistles, we show that our SRFormer achieves a 33.86dB PSNR score on the Urban100 dataset, which is 0.46dB higher than that of SwinIR but uses fewer parameters and computations. We hope our simple and effective approach can serve as a useful tool for future research in super-resolution model design. Our code is available at https://github.com/HVision-NKU/SRFormer.
****
Deep Homography Mixture for Single Image Rolling Shutter Correction
Weilong Yan, Robby T. Tan, Bing Zeng, Shuaicheng Liu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9868-9877
We present a deep homography mixture motion model for single image rolling shutter correction. Rolling shutter (RS) effects are often caused by row-wise exposure delay in the widely adopted CMOS sensor. Previous methods often require more than one frame for the correction, leading to data quality requirements. Few approaches address the more challenging task of single image RS correction, which often adopt designs like trajectory estimation or long rectangular kernels, to learn the camera motion parameters of an RS image, to restore the global shutter (GS) image. In this work, we adopt a more straightforward method to learn deep homography mixture motion between an RS image and its corresponding GS image, without large solution space or strict restrictions on image features. We show that dividing an image into blocks with a Gaussian weight of block scanlines fits well for the RS setting. Moreover, instead of directly learning the motion mapping, we learn coefficients that assemble several motion bases to produce the correction motion, where these bases are learned from the consecutive frames of natural videos beforehand. Experiments show that our method outperforms existing single RS methods statistically and visually, in both synthesized and real RS images.
****
Audio-Visual Glance Network for Efficient Video Recognition
Muhammad Adi Nugroho, Sangmin Woo, Sumin Lee, Changick Kim; Proceedings of the I

EEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10150-1015
9
Deep learning has made significant strides in video understanding tasks, but the computation required to classify lengthy and massive videos using clip-level video classifiers remains impractical and prohibitively expensive.
To address this issue, we propose Audio-Visual Glance Network (AVGN), which leverages the commonly available audio and visual modalities to efficiently process the spatio-temporally important parts of a video. AVGN firstly divides the video into snippets of image-audio clip pair and employs lightweight unimodal encoders to extract global visual features and audio features. To identify the important temporal segments, we use an Audio-Visual Temporal Saliency Transformer (AV-TeST) that estimates the saliency scores of each frame. To further increase efficiency in the spatial dimension, AVGN processes only the important patches instead of the whole images. We use an Audio-Enhanced Spatial Patch Attention (AESPA) module to produce a set of enhanced coarse visual features, which are fed to a policy network that produces the coordinates of the important patches. This approach enables us to focus only on the most important spatio-temporally parts of the video, leading to more efficient video recognition. Moreover, we incorporate various training techniques and multi-modal feature fusion to enhance the robustness and effectiveness of our AVGN. By combining these strategies, our AVGN sets new state-of-the-art performance in multiple video recognition benchmarks while achieving faster processing speed.
****
CLNeRF: Continual Learning Meets NeRF
Zhipeng Cai, Matthias Müller; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23185-23194
Novel view synthesis aims to render unseen views given a set of calibrated images. In practical applications, the coverage, appearance or geometry of the scene may change over time, with new images continuously being captured. Efficiently incorporating such continuous change is an open challenge. Standard NeRF benchmarks only involve scene coverage expansion. To study other practical scene changes, we propose a new dataset, World Across Time (WAT), consisting of scenes that change in appearance and geometry over time. We also propose a simple yet effective method, CLNeRF, which introduces continual learning (CL) to Neural Radiance Fields (NeRFs). CLNeRF combines generative replay and the Instant Neural Graphics Primitives (NGP) architecture to effectively prevent catastrophic forgetting and efficiently update the model when new data arrives. We also add trainable appearance and geometry embeddings to NGP, allowing a single compact model to handle complex scene changes. Without the need to store historical images, CLNeRF trained sequentially over multiple scans of a changing scene performs on-par with the upper bound model trained on all scans at once. Compared to other CL baselines CLNeRF performs much better across standard benchmarks and WAT. The source code, a demo, and the WAT dataset are available at https://github.com/IntelLabs/CLNeRF.
****
Rendering Humans from Object-Occluded Monocular Videos
Tiange Xiang, Adam Sun, Jiajun Wu, Ehsan Adeli, Li Fei-Fei; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3239-3250
3D understanding and rendering of moving humans from monocular videos is a challenging task. Although recent progress has enabled this task to some extent, it is still difficult to guarantee satisfactory results in real-world scenarios, where obstacles may block the camera view and cause partial occlusions in the captured videos. Existing methods cannot handle such defects due to two reasons. Firstly, the standard rendering strategy relies on point-point mapping, which could lead to dramatic disparities between the visible and occluded areas of the body. Secondly, the naive direct regression approach does not consider any feasibility criteria (i.e., prior information) for rendering under occlusions. To tackle the above drawbacks, we present OccNeRF, a neural rendering method that achieves better rendering of humans in severely occluded scenes. As direct solutions to the two drawbacks, we propose surface-based rendering by integrating geometry and

visibility priors. We validate our method on both simulated and real-world occlusions and demonstrate our method's superiority.
****
CrossMatch: Source-Free Domain Adaptive Semantic Segmentation via Cross-Modal Consistency Training
Yifang Yin, Wenmiao Hu, Zhenguang Liu, Guanfeng Wang, Shili Xiang, Roger Zimmermann; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21786-21796
Source-free domain adaptive semantic segmentation has gained increasing attention recently. It eases the requirement of full data access to the source domain by transferring knowledge only from a well-trained source model. However, reducing the uncertainty of the target pseudo labels becomes inevitably more challenging without the supervision of the labeled source data. In this work, we propose a novel asymmetric two-stream architecture that learns more robustly from noisy pseudo labels. Our approach simultaneously conducts dual-head pseudo label denoising and cross-modal consistency regularization. Towards the former, we introduce a multimodal auxiliary network during training (and discard it during inference), which effectively enhances the pseudo labels' correctness by leveraging the guidance from the depth information. Towards the latter, we enforce a new cross-modal pixel-wise consistency between the predictions of the two streams, encouraging our model to behave smoothly for both modality variance and image perturbations. It serves as an effective regularization to further reduce the impact of the inaccurate pseudo labels in source-free unsupervised domain adaptation. Experiments on GTA5 to Cityscapes and SYNTHIA to Cityscapes benchmarks demonstrate the superiority of our proposed method, obtaining the new state-of-the-art mIoU of 57.7% and 57.5%, respectively.
****
Out-of-Distribution Detection for Monocular Depth Estimation
Julia Hornauer, Adrian Holzbock, Vasileios Belagiannis; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 1911-1921
In monocular depth estimation, uncertainty estimation approaches mainly target the data uncertainty introduced by image noise. In contrast to prior work, we address the uncertainty due to lack of knowledge, which is relevant for the detection of data not represented by the training distribution, the so-called out-of-distribution (OOD) data. Motivated by anomaly detection, we propose to detect OOD images from an encoder-decoder depth estimation model based on the reconstruction error. Given the features extracted with the fixed depth encoder, we train an image decoder for image reconstruction using only in-distribution data. Consequently, OOD images result in a high reconstruction error, which we use to distinguish between in- and out-of-distribution samples. We built our experiments on the standard NYU Depth V2 and KITTI benchmarks as in-distribution data. Our post hoc method performs astonishingly well on different models and outperforms existing uncertainty estimation approaches without modifying the trained encoder-decoder depth estimation model.
****
STEPs: Self-Supervised Key Step Extraction and Localization from Unlabeled Procedural Videos
Anshul Shah, Benjamin Lundell, Harpreet Sawhney, Rama Chellappa; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 10375-10387
We address the problem of extracting key steps from unlabeled procedural videos, motivated by the potential of Augmented Reality (AR) headsets to revolutionize job training and performance. We decompose the problem into two steps: representation learning and key steps extraction. We propose a training objective, Bootstrapped Multi-Cue Contrastive (BMC2) loss to learn discriminative representations for various steps without any labels. Different from prior works, we develop techniques to train a light-weight temporal module which uses off-the-shelf features for self supervision. Our approach can seamlessly leverage information from multiple cues like optical flow, depth or gaze to learn discriminative features for key-steps, making it amenable for AR applications. We finally extract key ste

ps via a tunable algorithm that clusters the representations and samples. We sho
w significant improvements over prior works for the task of key step localizatio
n and phase classification. Qualitative results demonstrate that the extracted k
ey steps are meaningful and succinctly represent various steps of the procedural
 tasks.
****
Improving Equivariance in State-of-the-Art Supervised Depth and Normal Predictor
s
Yuanyi Zhong, Anand Bhattad, Yu-Xiong Wang, David Forsyth; Proceedings of the IE
EE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21775-21785
Dense depth and surface normal predictors should possess the equivariant propert
y to cropping-and-resizing -- cropping the input image should result in cropping
 the same output image. However, we find that state-of-the-art depth and normal
predictors, despite having strong performances, surprisingly do not respect equi
variance. The problem exists even when crop-and-resize data augmentation is empl
oyed during training. To remedy this, we propose an equivariant regularization t
echnique, consisting of an averaging procedure and a self-consistency loss, to e
xplicitly promote cropping-and-resizing equivariance in depth and normal network
s. Our approach can be applied to both CNN and Transformer architectures, does n
ot incur extra cost during testing, and notably improves the supervised and semi
-supervised learning performance of dense predictors on Taskonomy tasks. Finally
, finetuning with our loss on unlabeled images improves not only equivariance bu
t also accuracy of state-of-the-art depth and normal predictors when evaluated o
n NYU-v2.
****
Towards Robust and Smooth 3D Multi-Person Pose Estimation from Monocular Videos
in the Wild
Sungchan Park, Eunyi You, Inhoe Lee, Joonseok Lee; Proceedings of the IEEE/CVF I
nternational Conference on Computer Vision (ICCV), 2023, pp. 14772-14782
3D pose estimation is an invaluable task in computer vision with various practic
al applications. Especially, 3D pose estimation for multi-person from a monocula
r video (3DMPPE) is particularly challenging and is still largely uncharted, far
 from applying to in-the-wild scenarios yet. We pose three unresolved issues wit
h the existing methods: lack of robustness on unseen views during training, vuln
erability to occlusion, and severe jittering in the output. As a remedy, we prop
ose POTR-3D, the first realization of a sequence-to-sequence 2D-to-3D lifting mo
del for 3DMPPE, powered by a novel geometry-aware data augmentation strategy, ca
pable of generating unbounded data with a variety of views while caring about th
e ground plane and occlusions. Through extensive experiments, we verify that the
 proposed model and data augmentation robustly generalizes to diverse unseen vie
ws, robustly recovers the poses against heavy occlusions, and reliably generates
 more natural and smoother outputs. The effectiveness of our approach is verifie
d not only by achieving the state-of-the-art performance on public benchmarks, b
ut also by qualitative results on more challenging in-the-wild videos. Demo vide
os are available at https://www.youtube.com/@potr3d.
****
Reducing Training Time in Cross-Silo Federated Learning Using Multigraph Topolog
y
Tuong Do, Binh X. Nguyen, Vuong Pham, Toan Tran, Erman Tjiputra, Quang D. Tran,
Anh Nguyen; Proceedings of the IEEE/CVF International Conference on Computer Vis
ion (ICCV), 2023, pp. 19409-19419
Federated learning is an active research topic since it enables several particip
ants to jointly train a model without sharing local data. Currently, cross-silo
federated learning is a popular training setting that utilizes a few hundred rel
iable data silos with high-speed access links to training a model. While this ap
proach has been widely applied in real-world scenarios, designing a robust topol
ogy to reduce the training time remains an open problem. In this paper, we prese
nt a new multigraph topology for cross-silo federated learning. We first constru
ct the multigraph using the overlay graph. We then parse this multigraph into di
fferent simple graphs with isolated nodes. The existence of isolated nodes allow

s us to perform model aggregation without waiting for other nodes, hence effectively reducing the training time. Intensive experiments on three public datasets show that our proposed method significantly reduces the training time compared with recent state-of-the-art topologies while maintaining the accuracy of the learned model.
****
Counting Crowds in Bad Weather
Zhi-Kai Huang, Wei-Ting Chen, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23308-23319
Crowd counting has recently attracted significant attention in the field of computer vision due to its wide applications to image understanding. Numerous methods have been proposed and achieved state-of-the-art performance for real-world tasks. However, existing approaches do not perform well under adverse weather such as haze, rain, and snow since the visual appearances of crowds in such scenes are drastically different from those images in clear weather of typical datasets. In this paper, we propose a method for robust crowd counting in adverse weather scenarios. Instead of using a two-stage approach that involves image restoration and crowd counting modules, our model learns effective features and adaptive queries to account for large appearance variations. With these weather queries, the proposed model can learn the weather information according to the degradation of the input image and optimize with the crowd counting module simultaneously. Experimental results show that the proposed algorithm is effective in counting crowds under different weather types on benchmark datasets. The source code and trained models will be made available to the public.
****
FreeDoM: Training-Free Energy-Guided Conditional Diffusion Model
Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, Jian Zhang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 23174-23184
Recently, conditional diffusion models have gained popularity in numerous applications due to their exceptional generation ability. However, many existing methods are training-required. They need to train a time-dependent classifier or a condition-dependent score estimator, which increases the cost of constructing conditional diffusion models and is inconvenient to transfer across different conditions. Some current works aim to overcome this limitation by proposing training-free solutions, but most can only be applied to a specific category of tasks and not to more general conditions. In this work, we propose a training-Free conditional Diffusion Model (FreeDoM) used for various conditions. Specifically, we leverage off-the-shelf pre-trained networks, such as a face detection model, to construct time-independent energy functions, which guide the generation process without requiring training. Furthermore, because the construction of the energy function is very flexible and adaptable to various conditions, our proposed FreeDoM has a broader range of applications than existing training-free methods. FreeDoM is advantageous in its simplicity, effectiveness, and low cost. Experiments demonstrate that FreeDoM is effective for various conditions and suitable for diffusion models of diverse data domains, including image and latent code domains.
****
UniT3D: A Unified Transformer for 3D Dense Captioning and Visual Grounding
Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, Angel X. Chang; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18109-18119
Performing 3D dense captioning and visual grounding requires a common and shared understanding of the underlying multimodal relationships. However, despite some previous attempts on connecting these two related tasks with highly task-specific neural modules, it remains understudied how to explicitly depict their shared nature to learn them simultaneously. In this work, we propose UniT3D, a simple yet effective fully unified transformer-based architecture for jointly solving 3D visual grounding and dense captioning. UniT3D enables learning a strong multimodal representation across the two tasks through a supervised joint pre-training

scheme with bidirectional and seq-to-seq objectives. With a generic architecture design, UniT3D allows expanding the pre-training scope to more various training sources such as the synthesized data from 2D prior knowledge to benefit 3D vision-language tasks. Extensive experiments and analysis demonstrate that UniT3D obtains significant gains for 3D dense captioning and visual grounding.
****
SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition
Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, Sebastien Ourselin; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 21074-21084
This paper introduces SKiT, a fast Key information Transformer for phase recognition of videos. Unlike previous methods that rely on complex models to capture long-term temporal information, SKiT accurately recognizes high-level stages of videos using an efficient key pooling operation. This operation records important key information by retaining the maximum value recorded from the beginning up to the current video frame, with a time complexity of O(1). Experimental results on Cholec80 and AutoLaparo surgical datasets demonstrate the ability of our model to recognize phases in an online manner. SKiT achieves higher performance than state-of-the-art methods with an accuracy of 92.5% and 82.9% on Cholec80 and AutoLaparo, respectively, while running the temporal model eight times faster (7ms v.s. 55ms) than LoViT, which uses ProbSparse to capture global information. We highlight that the inference time of SKiT is constant, and independent from the input length, making it a stable choice for keeping a record of important global information, that appears on long surgical videos, essential for phase recognition. To sum up, we propose an effective and efficient model for surgical phase recognition that leverages key global information. This has an intrinsic value when performing this task in an online manner on long surgical videos for stable real-time surgical recognition systems.
****
Clustering based Point Cloud Representation Learning for 3D Analysis
Tuo Feng, Wenguan Wang, Xiaohan Wang, Yi Yang, Qinghua Zheng; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8283-8294
Point cloud analysis (such as 3D segmentation and detection) is a challenging task, because of not only the irregular geometries of many millions of unordered points, but also the great variations caused by depth, viewpoint, occlusion, etc. Current studies put much focus on the adaption of neural networks to the complex geometries of point clouds, but are blind to a fundamental question: how to learn an appropriate point embedding space that is aware of both discriminative semantics and challenging variations? As a response, we propose a clustering based supervised learning scheme for point cloud analysis. Unlike current de-facto, scene-wise training paradigm, our algorithm conducts within-class clustering on the point embedding space for automatically discovering subclass patterns which are latent yet representative across scenes. The mined patterns are, in turn, used to repaint the embedding space, so as to respect the underlying distribution of the entire training dataset and improve the robustness to the variations. Our algorithm is principled and readily pluggable to modern point cloud segmentation networks during training, without extra overhead during testing. With various 3D network architectures (i.e., voxel-based, point-based, Transformer-based, automatically searched), our algorithm shows notable improvements on famous point cloud segmentation datasets (i.e., 2.0-2.6% on single-scan and 2.0-2.2% multi-scan of SemanticKITTI, 1.8-1.9% on S3DIS, in terms of mIoU). Our algorithm also demonstrates utility in 3D detection, showing 2.0-3.4% mAP gains on KITTI. Our code is released at: https://github.com/FengZicai/Cluster3Dseg/.
****
Automatic Network Pruning via Hilbert-Schmidt Independence Criterion Lasso under Information Bottleneck Principle
Song Guo, Lei Zhang, Xiawu Zheng, Yan Wang, Yuchao Li, Fei Chao, Chenglin Wu, Shengchuan Zhang, Rongrong Ji; Proceedings of the IEEE/CVF International Conferenc

e on Computer Vision (ICCV), 2023, pp. 17458-17469
Most existing neural network pruning methods hand-crafted their importance crite ria and structures to prune. This constructs heavy and unintended dependencies o n heuristics and expert experience for both the objective and the parameters of the pruning approach. In this paper, we try to solve this problem by introducing a principled and unified framework based on Information Bottleneck (IB) theory, which further guides us to an automatic pruning approach. Specifically, we firs t formulate the channel pruning problem from an IB perspective, and then impleme nt the IB principle by solving a Hilbert-Schmidt Independence Criterion (HSIC) L asso problem under certain conditions. Based on the theoretical guidance, we the n provide an automatic pruning scheme by searching for global penalty coefficien ts. Verified by extensive experiments, our method yields state-of-the-art perfor mance on various benchmark networks and datasets. For example, with VGG-16, we a chieve a 60%-FLOPs reduction by removing 76% of the parameters, with an improvem ent of 0.40% in top-1 accuracy on CIFAR-10. With ResNet-50, we achieve a 56%-FLO Ps reduction by removing 50% of the parameters, with a small loss of 0.08% in th e top-1 accuracy on ImageNet. The code is available at https://github.com/sunggo /APIB.
****
Forecast-MAE: Self-supervised Pre-training for Motion Forecasting with Masked Au toencoders
Jie Cheng, Xiaodong Mei, Ming Liu; Proceedings of the IEEE/CVF International Con ference on Computer Vision (ICCV), 2023, pp. 8679-8689
This study explores the application of self-supervised learning (SSL) to the tas k of motion forecasting, an area that has not yet been extensively investigated despite the widespread success of SSL in computer vision and natural language pr ocessing. To address this gap, we introduce Forecast-MAE, an extension of the ma sk autoencoders framework that is specifically designed for self-supervised lear ning of the motion forecasting task. Our approach includes a novel masking strat egy that leverages the strong interconnections between agents' trajectories and road networks, involving complementary masking of agents' future or history traj ectories and random masking of lane segments. Our experiments on the challenging Argoverse 2 motion forecasting benchmark show that Forecast-MAE, which utilizes standard Transformer blocks with minimal inductive bias, achieves competitive p erformance compared to state-of-the-art methods that rely on supervised learning and sophisticated designs. Moreover, it outperforms the previous self-supervise d learning method by a significant margin. Code is available at https://github.c om/jchengai/forecast-mae.
****
Efficient Transformer-based 3D Object Detection with Dynamic Token Halting
Mao Ye, Gregory P. Meyer, Yuning Chai, Qiang Liu; Proceedings of the IEEE/CVF In ternational Conference on Computer Vision (ICCV), 2023, pp. 8438-8450
Balancing efficiency and accuracy is a long-standing problem for deploying deep learning models. The trade-off is even more important for real-time safety-criti cal systems like autonomous vehicles. In this paper, we propose an effective app roach for accelerating transformer-based 3D object detectors by dynamically halt ing tokens at different layers depending on their contribution to the detection task. Although halting a token is a non-differentiable operation, our method all ows for differentiable end-to-end learning by leveraging an equivalent different iable forward-pass. Furthermore, our framework allows halted tokens to be reused to inform the model's predictions through a straightforward token recycling mec hanism. Our method significantly improves the Pareto frontier of efficiency vers us accuracy when compared with the existing approaches. By halting tokens and in creasing model capacity, we are able to improve the baseline model's performance without increasing the model's latency on the Waymo Open Dataset.
****
Neglected Free Lunch - Learning Image Classifiers Using Annotation Byproducts
Dongyoon Han, Junsuk Choe, Seonghyeok Chun, John Joon Young Chung, Minsuk Chang, Sangdoo Yun, Jean Y. Song, Seong Joon Oh; Proceedings of the IEEE/CVF Internati onal Conference on Computer Vision (ICCV), 2023, pp. 20200-20212

Supervised learning of image classifiers distills human knowledge into a paramet ric model through pairs of images and corresponding labels (X,Y). We argue that this simple and widely used representation of human knowledge neglects rich auxi liary information from the annotation procedure, such as the time-series of mous e traces and clicks left after image selection. Our insight is that such annotat ion byproducts Z provide approximate human attention that weakly guides the mode l to focus on the foreground cues, reducing spurious correlations and discouragi ng shortcut learning. To verify this, we create ImageNet-AB and COCO-AB. They ar e ImageNet and COCO training sets enriched with sample-wise annotation byproduct s, collected by replicating the respective original annotation tasks. We refer t o the new paradigm of training models with annotation byproducts as learning usi ng annotation byproducts (LUAB). We show that a simple multitask loss for regres sing Z together with Y already improves the generalisability and robustness of t he learned models. Compared to the original supervised learning, LUAB does not r equire extra annotation costs. ImageNet-AB and COCO-AB are at https://github.com /naver-ai/NeglectedFreeLunch.
****
Rethinking the Role of Pre-Trained Networks in Source-Free Domain Adaptation
Wenyu Zhang, Li Shen, Chuan-Sheng Foo; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18841-18851
Source-free domain adaptation (SFDA) aims to adapt a source model trained on a f ully-labeled source domain to an unlabeled target domain. Large-data pre-trained networks are used to initialize source models during source training, and subse quently discarded. However, source training can cause the model to overfit to so urce data distribution and lose applicable target domain knowledge. We propose t o integrate the pre-trained network into the target adaptation process as it has diversified features important for generalization and provides an alternate vie w of features and classification decisions different from the source model. We p ropose to distil useful target domain information through a co-learning strategy to improve target pseudolabel quality for finetuning the source model. Evaluati on on 4 benchmark datasets show that our proposed strategy improves adaptation p erformance and can be successfully integrated with existing SFDA methods. Levera ging modern pre-trained networks that have stronger representation learning abil ity in the co-learning strategy further boosts performance.
****
RLIPv2: Fast Scaling of Relational Language-Image Pre-Training
Hangjie Yuan, Shiwei Zhang, Xiang Wang, Samuel Albanie, Yining Pan, Tao Feng, Ji anwen Jiang, Dong Ni, Yingya Zhang, Deli Zhao; Proceedings of the IEEE/CVF Inter national Conference on Computer Vision (ICCV), 2023, pp. 21649-21661
Relational Language-Image Pre-training (RLIP) aims to align vision representatio ns with relational texts, thereby advancing the capability of relational reasoni ng in computer vision tasks. However, hindered by the slow convergence of RLIPv1 architecture and the limited availability of existing scene graph data, scaling RLIPv1 is challenging. In this paper, we propose RLIPv2, a fast converging mode l that enables the scaling of relational pre-training to large-scale pseudo-labe lled scene graph data. To enable fast scaling, RLIPv2 introduces Asymmetric Lang uage-Image Fusion (ALIF), a mechanism that facilitates earlier and deeper gated cross-modal fusion with sparsified language encoding layers. ALIF leads to compa rable or better performance than RLIPv1 in a fraction of the time for pre-traini ng and fine-tuning. To obtain scene graph data at scale, we extend object detect ion datasets with free-form relation labels by introducing a captioner (e.g., BL IP) and a designed Relation Tagger. The Relation Tagger assigns BLIP-generated r elation texts to region pairs, thus enabling larger-scale relational pre-trainin g. Through extensive experiments conducted on Human-Object Interaction Detection and Scene Graph Generation, RLIPv2 shows state-of-the-art performance on three benchmarks under fully-finetuning, few-shot and zero-shot settings. Notably, the largest RLIPv2 achieves 23.29mAP on HICO-DET without any fine-tuning, yields 32 .22mAP with just 1% data and yields 45.09mAP with 100% data. Code and models are publicly available at https://github.com/JacobYuan7/RLIPv2.
****

TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective
Jun Dan, Yang Liu, Haoyu Xie, Jiankang Deng, Haoran Xie, Xuansong Xie, Baigui Sun; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20642-20653
Vision Transformers (ViTs) have demonstrated powerful representation ability in various visual tasks thanks to their intrinsic data-hungry nature. However, we unexpectedly find that ViTs perform vulnerably when applied to face recognition (FR) scenarios with extremely large datasets. We investigate the reasons for this phenomenon and discover that the existing data augmentation approach and hard sample mining strategy are incompatible with ViTs-based FR backbone due to the lack of tailored consideration on preserving face structural information and leveraging each local token information. To remedy these problems, this paper proposes a superior FR model called TransFace, which employs a patch-level data augmentation strategy named DPAP and a hard sample mining strategy named EHSM. Specially, DPAP randomly perturbs the amplitude information of dominant patches to expand sample diversity, which effectively alleviates the overfitting problem in ViTs. EHSM utilizes the information entropy in the local tokens to dynamically adjust the importance weight of easy and hard samples during training, leading to a more stable prediction. Experiments on several benchmarks demonstrate the superiority of our TransFace. Code and models are available at https://github.com/DanJun6737/TransFace.
****
LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models
Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, Yu Su; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 2998-3009
This study focuses on using large language models (LLMs) as a planner for embodied agents that can follow natural language instructions to complete complex tasks in a visually-perceived environment. The high data cost and poor sample efficiency of existing methods hinders the development of versatile agents that are capable of many tasks and can learn new tasks quickly. In this work, we propose a novel method, LLM-Planner, that harnesses the power of large language models to do few-shot planning for embodied agents. We further propose a simple but effective way to enhance LLMs with physical grounding to generate and update plans that are grounded in the current environment. Experiments on the ALFRED dataset show that our method can achieve very competitive few-shot performance: Despite using less than 0.5% of paired training data, LLM-Planner achieves competitive performance with recent baselines that are trained using the full training data. Existing methods can barely complete any task successfully under the same few-shot setting.
Our work opens the door for developing versatile and sample-efficient embodied agents that can quickly learn many tasks.
****
Exploring Model Transferability through the Lens of Potential Energy
Xiaotong Li, Zixuan Hu, Yixiao Ge, Ying Shan, Ling-Yu Duan; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 5429-5438
Transfer learning has become crucial in computer vision tasks due to the vast availability of pre-trained deep learning models. However, selecting the optimal pre-trained model from a diverse pool for a specific downstream task remains a challenge. Existing methods for measuring the transferability of pre-trained models rely on statistical correlations between encoded static features and task labels, but they overlook the impact of underlying representation dynamics during fine-tuning, leading to unreliable results, especially for self-supervised models. In this paper, we present an insightful physics-inspired approach named PED to address these challenges. We reframe the challenge of model selection through the lens of potential energy and directly model the interaction forces that influence fine-tuning dynamics. By capturing the motion of dynamic representations to decline the potential energy within a force-driven physical model, we can acquir

e an enhanced and more stable observation for estimating transferability. The experimental results on 10 downstream tasks and 12 self-supervised models demonstrate that our approach can seamlessly integrate into existing ranking techniques and enhance their performances, revealing its effectiveness for the model selection task and its potential for understanding the mechanism in transfer learning. Code is available at https://github.com/lixiaotong97/PED.
****
Video Task Decathlon: Unifying Image and Video Tasks in Autonomous Driving
Thomas E. Huang, Yifan Liu, Luc Van Gool, Fisher Yu; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 8647-8657
Performing multiple heterogeneous visual tasks in dynamic scenes is a hallmark of human perception capability. Despite remarkable progress in image and video recognition via representation learning, current research still focuses on designing specialized networks for singular, homogeneous, or simple combination of tasks. We instead explore the construction of a unified model for major image and video recognition tasks in autonomous driving with diverse input and output structures. To enable such an investigation, we design a new challenge, Video Task Decathlon (VTD), which includes ten representative image and video tasks spanning classification, segmentation, localization, and association of objects and pixels. On VTD, we develop our unified network, VTDNet, that uses a single structure and a single set of weights for all ten tasks. VTDNet groups similar tasks and employs task interaction stages to exchange information within and between task groups. Given the impracticality of labeling all tasks on all frames, and the performance degradation associated with joint training of many tasks, we design a Curriculum training, Pseudo-labeling, and Fine-tuning (CPF) scheme to successfully train VTDNet on all tasks and mitigate performance loss. Armed with CPF, VTDNet significantly outperforms its single-task counterparts on most tasks with only 20% overall computations. VTD is a promising new direction for exploring the unification of perception tasks in autonomous driving.
****
Aria Digital Twin: A New Benchmark Dataset for Egocentric 3D Machine Perception
Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, Yuheng (Carl) Ren; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 20133-20143
We introduce the Aria Digital Twin (ADT) - an egocentric dataset captured using Aria glasses with extensive object, environment, and human level ground truth. This ADT release contains 200 sequences of real-world activities conducted by Aria wearers in two real indoor scenes with 398 object instances (344 stationary and 74 dynamic). Each sequence consists of: a) raw data of two monochrome camera streams, one RGB camera stream, two IMU streams; b) complete sensor calibration; c) ground truth data including continuous 6-degree-of-freedom (6DoF) poses of the Aria devices, object 6DoF poses, 3D eye gaze vectors, 3D human poses, 2D image segmentations, image depth maps; and d) photo-realistic synthetic renderings. To the best of our knowledge, there is no existing egocentric dataset with a level of accuracy, photo-realism and comprehensiveness comparable to ADT. By contributing ADT to the research community, our mission is to set a new standard for evaluation in the egocentric machine perception domain, which includes very challenging research problems such as 3D object detection and tracking, scene reconstruction and understanding, sim-to-real learning, human pose prediction - while also inspiring new machine perception tasks for augmented reality (AR) applications. To kick start exploration of the ADT research use cases, we evaluated several existing state-of-the-art methods for object detection, segmentation and image translation tasks that demonstrate the usefulness of ADT as a benchmarking dataset.
****
PreSTU: Pre-Training for Scene-Text Understanding
Jihyung Kil, Soravit Changpinyo, Xi Chen, Hexiang Hu, Sebastian Goodman, Wei-Lun Chao, Radu Soricut; Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15270-15280
The ability to recognize and reason about text embedded in visual inputs is ofte

n lacking in vision-and-language (V&L) models, perhaps because V&L pre-training
methods have often failed to include such an ability in their training objective
. In this paper, we propose PreSTU, a novel pre-training recipe dedicated to sce
ne-text understanding (STU). PreSTU introduces OCR-aware pre-training objectives
 that encourage the model to recognize text from an image and connect it to the
rest of the image content. We implement PreSTU using a simple transformer-based
encoder-decoder architecture, combined with large-scale image-text datasets with
 scene text obtained from an off-the-shelf OCR system. We empirically demonstrat
e the effectiveness of this pre-training approach on eight visual question answe
ring and four image captioning benchmarks.
****